PageRank

Varun Kumar and Marco Groenendaal

October 5, 2015

Table of Contents

Background

Search engine

Searches for information on World Wide Web based on keywords

Programmed to rank sites based on popularity, relevance

Three functions:

- Crawling
- Indexing
- Searching

Source: Brin and Page, 1998

History

Archie (1990) - first search engine

Hosted an index of directory listings (no contents)

WebCrawler and Lycos (1994)

Able to search words of webpage

AskJeeves (1997)

Natural language search engine, ranked links by popularity

Yahoo! (pre-Google)

Searchable directory

Source: Wordstream, “Internet Search Engines History”

The problem

Text-based ranking systems of old search engines were easily manipulated

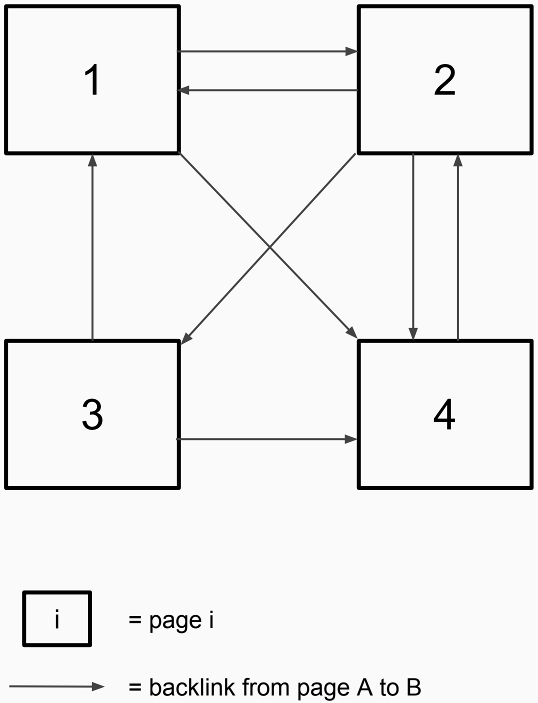
Ex: Searching "Internet" pre-Google

"Any evaluation strategy which counts replicable features of web pages is prone to manipulation" (Page)

"Pages [were] engineered to manipulate search engine ranking functions" (Page)

Source: Brin and Page, 1998

The Internet as a directed graph



Nodes = pages, edges = links

Backlinks: hyperlinks from a page to another page

i.e. edges directed towards nodes

Pages with many backlinks considered relevant

Source: Brin and Page, 1998

Linear algebra tutorial

Definitions and properties

Coefficient Matrix: An *m x n* matrix, **A**, containing only the coefficients of a linear system

Linear Independence: The columns of the matrix, **A**, are linearly independent if and only if the equation **Ax** = 0 has the trivial solution

Column-Stochastic Matrix: A square matrix, **A**, where every entry is non-negative [$\mathbf{A}_{ij} \geq 0$], the entries in each column sum to 1 [for j=1 to n, $\sum \mathbf{A}_{ij} = 1$], and has an λ = 1

Eigenvalue: a scalar, λ, such that **Ax = λx**, holds for some nonzero vector **x**

Eigenvector: a nonzero vector ***x*** of an *n x n* matrix, **A**, such that **Ax = λx.** Any nonzero multiple of an eigenvector is also an eigenvector.

Eigenspace: The null space of (**A** - λ)**I**

Source: Aerostudents, Bryan & Leise, 2006

Definitions and properties

Markov chain: A collection of random variables $\{\boldsymbol{X_t}\}$ (t > 0) with the property that given the present, the future is conditionally independent of the past. i.e.

$$P\,(X_t = j \mid X_{t-1} = i_{t-1}, X_{t-2} = i_{t-2}, \ldots, X_0 = i_0) = P\,(X_t = j \mid X_{t-1} = i_{t-1})$$

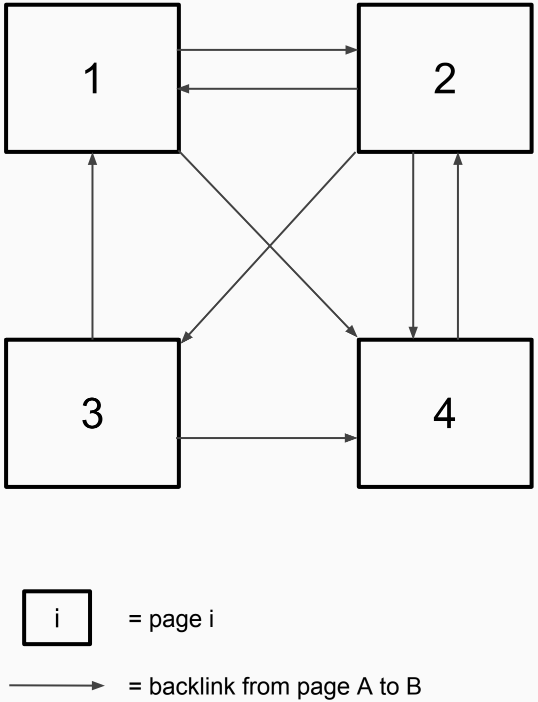
Column-Stochastic Matrix: A square matrix, **A**, where every entry is non-negative [$\mathbf{A}_{i\,j} \geq 0$], the entries in each column sum to 1 [for j=1 to n, $\sum \mathbf{A}_{i\,j} = 1$], and has an $\lambda = 1$

Random walk: A random process consisting of a sequence of discrete steps of a fixed length; an example of a Markov chain.

Source: Weisstein, Eric

PageRank basics & simple example

Example: web is n = 4 pages (1/6)



x_k = # of backlinks to page *k*

In this example:

$x_1 = 2$

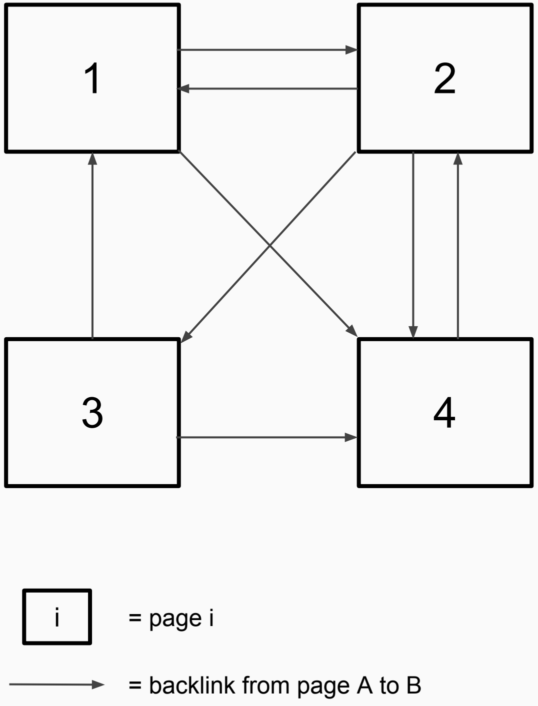
$x_2 = 2$

$x_3 = 1$ [least important]

$x_4 = 3$ [most important]

Source: Bryan & Leise, 2006

Example: web is n = 4 pages (2/6)



x_k = # of backlinks to page *k*

In this example:

$x_1 = 2$

$x_2 = 2$

$x_3 = 1$ [least important]

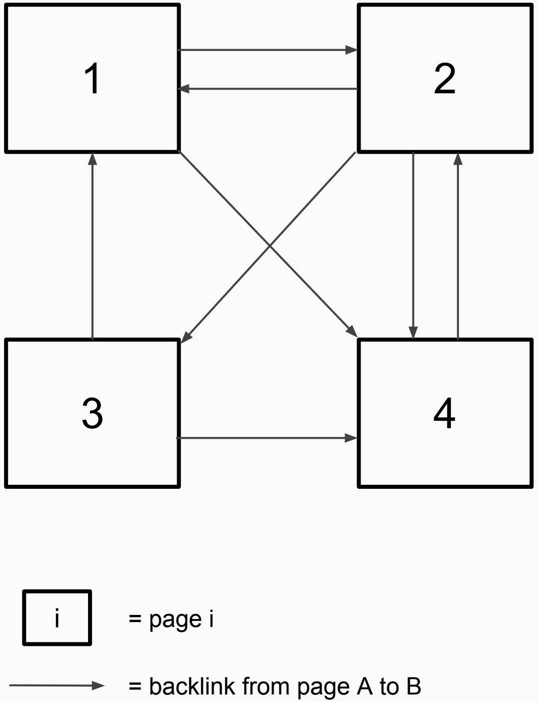
$x_4 = 3$ [most important]

PROBLEM:

- Are all links equal?
- What if a page has a large number of links?

Source: Bryan & Leise, 2006

Example: web is n = 4 pages (3/6)



ANSWER:

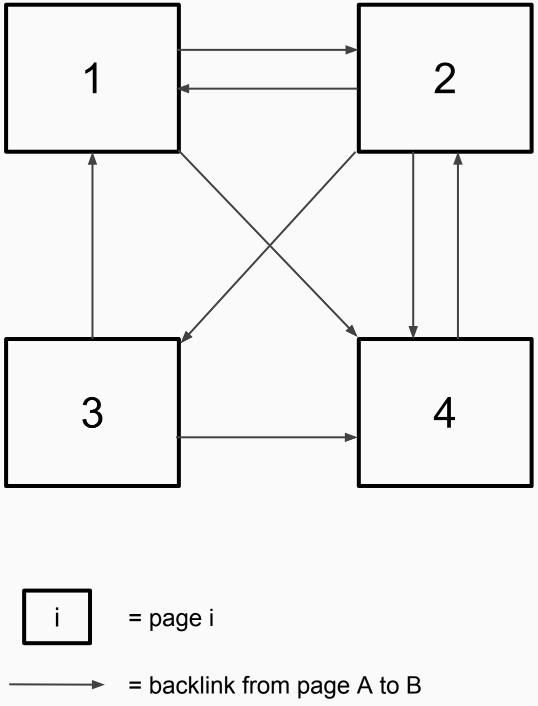
- Each page *j* gets a total vote of 1, that is weighted by page *j'* *s* score, and divided evenly by its outgoing links
- $x_k = \sum x_j / n_j$
  - x_j = page score of page *j*
  - n_j = # of links out of page *j*

In this example:

$$
\begin{aligned}
x_1 &= \phantom{x_1/2 + {}} x_2/3 + x_3/2 \\
x_2 &= x_1/2 \phantom{{} + x_2/3 + x_3/2} + x_4 \\
x_3 &= \phantom{x_1/2 + {}} x_2/3 \\
x_4 &= x_1/2 + x_2/3 + x_3/2
\end{aligned}
$$

Source: Bryan & Leise, 2006

Example: web is n = 4 pages (4/6)



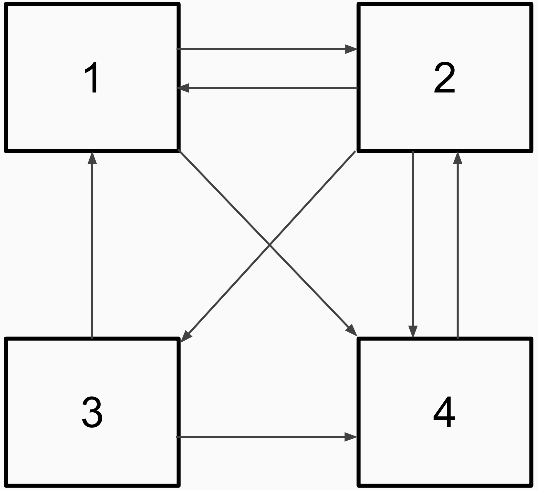
METHOD 1: We can rewrite as the link matrix, **A**, of coefficients and the unique, nonnegative eigenvector, **x**, with eigenvalues of 1.

Let

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} \quad \text{s.t. } \boldsymbol{x} = \mathbf{A}\boldsymbol{x} \quad \text{and } \mathbf{A} = \begin{pmatrix} 0 & \frac{1}{3} & \frac{1}{2} & 0 \\ \frac{1}{2} & 0 & 0 & 1 \\ 0 & \frac{1}{3} & 0 & 0 \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{2} & 0 \end{pmatrix} \text{ with } \lambda\text{'s} = 1$$

Source: Bryan & Leise, 2006

Example: web is n = 4 pages (5/6)



$\boxed{i}$ = page i

$\longrightarrow$ = backlink from page A to B

METHOD 1: We can rewrite as the link matrix, **A**, of coefficients and the unique, nonnegative eigenvector, **x**, with eigenvalues of 1.

Let

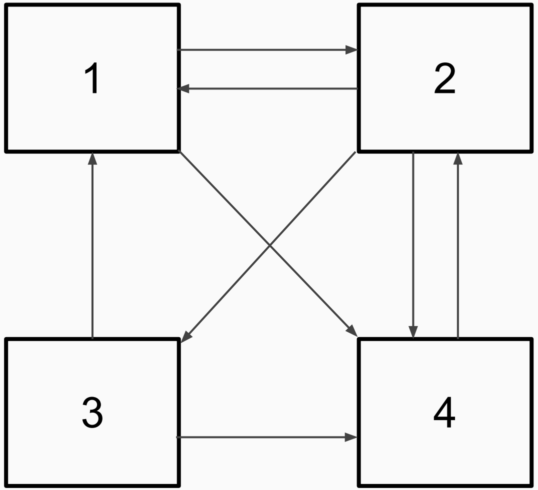
$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} \quad \text{s.t. } \boldsymbol{x} = \mathbf{A}\boldsymbol{x} \quad \text{and } \mathbf{A} = \begin{pmatrix} 0 & \frac{1}{3} & \frac{1}{2} & 0 \\ \frac{1}{2} & 0 & 0 & 1 \\ 0 & \frac{1}{3} & 0 & 0 \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{2} & 0 \end{pmatrix} \quad \text{with } \lambda\text{'s} = 1$$

By setting $\mathbf{A}\boldsymbol{x} = \lambda\boldsymbol{x}$ or $(\mathbf{A} - \lambda\, \mathbf{I})\boldsymbol{x} = \mathbf{0}$, we find that

$$\boldsymbol{x} = \begin{pmatrix} 2/3 \\ 4/3 \\ 4/9 \\ 1 \end{pmatrix} \underset{\times\, 9/31}{=} \begin{pmatrix} 0.194 \\ 0.387 \\ 0.129 \\ 0.290 \end{pmatrix} \quad \text{s.t. } \sum \boldsymbol{x} = 1 \;\;\&\;\; \dim[\mathbf{V}_1(\mathbf{A})] = 1$$

Source: Bryan & Leise, 2006

Example: web is n = 4 pages (6/6)



i = page i

→ = backlink from page A to B

METHOD 2: This can be seen as a random walk. **A** is the transition matrix, **x** is the vector of stationary probabilities, and *x* is the vector of initial probabilities (¼).

Let

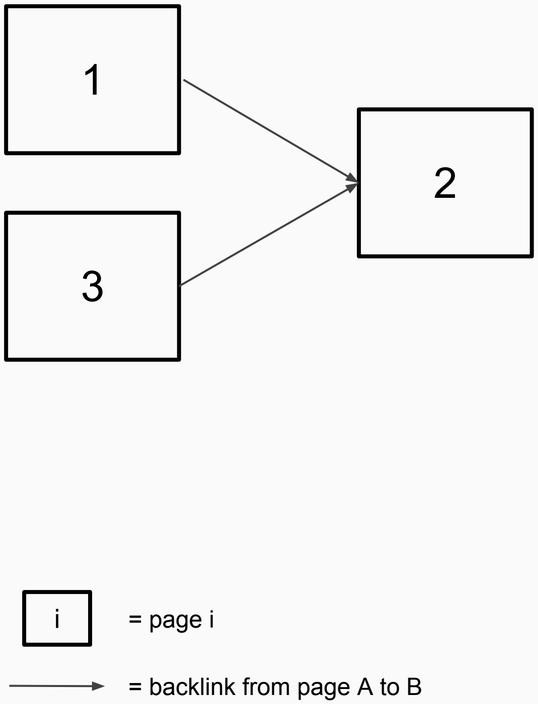
$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} \quad x = \begin{pmatrix} 1/4 \\ 1/4 \\ 1/4 \\ 1/4 \end{pmatrix} \quad \text{and } \mathbf{A} = \begin{pmatrix} 0 & 1/3 & 1/2 & 0 \\ 1/2 & 0 & 0 & 1 \\ 0 & 1/3 & 0 & 0 \\ 1/2 & 1/3 & 1/2 & 0 \end{pmatrix}$$

To solve $\mathbf{A}^{\infty}x$, we solve for the stationary distribution through $x^T = x^T\mathbf{A}$.

Source: Bryan & Leise, 2006

Issues & complex example

Issues with simple model - dangling node (1/2)

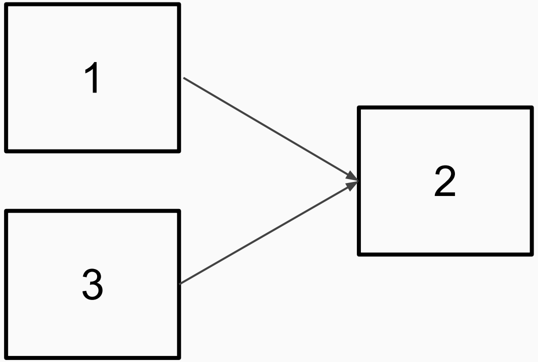


Using prior method, the calculated rank of every page is zero

Obviously not true, since page 2 is most important

Page 2 is called a **dangling node**

Issues with simple model - dangling node (2/2)



Using prior method, the calculated rank of every page is zero

Obviously not true, since page 2 is most important

Page 2 is called a **dangling node**

Dangling nodes

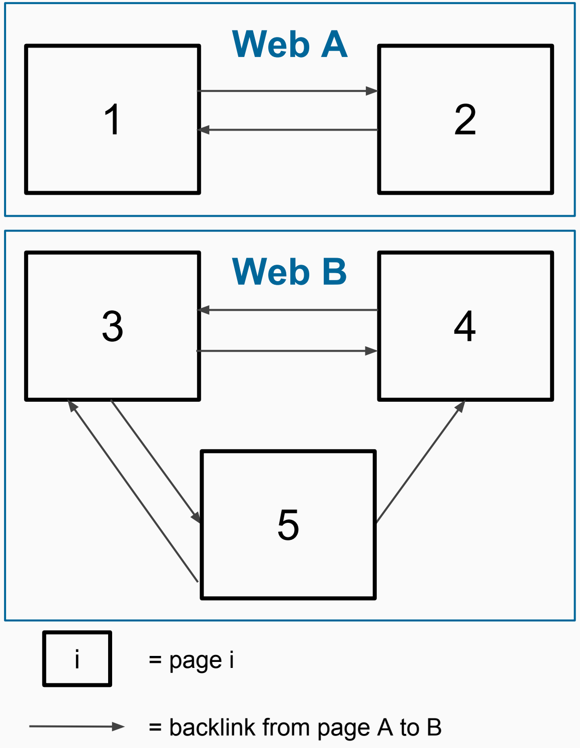
- A is **column-substochastic**
  - Entries in each column sum to less than or equal to 1 [for j=1 to n, $\sum \mathbf{A}_{ij} \leq 1$]
  - $\lambda \leq 1$

i = page i

= backlink from page A to B

Source: Bryan & Leise, 2006

Issues with simple model - disconnected graphs (1/2)

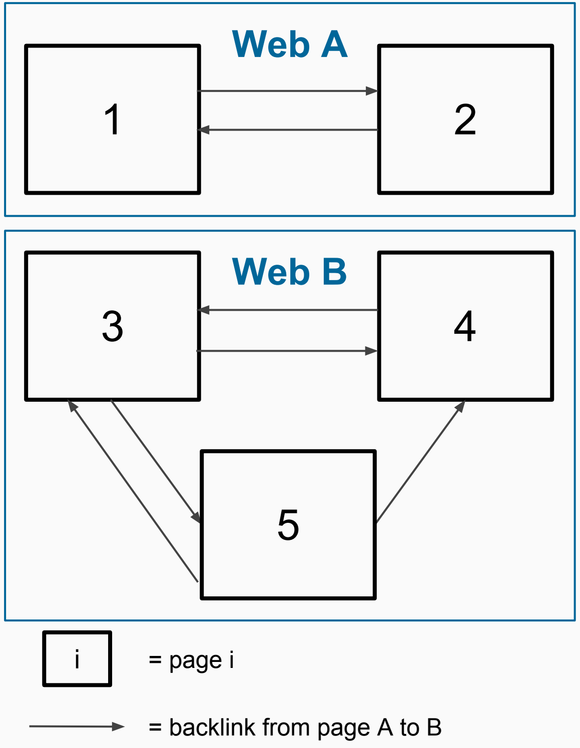


We let

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix} \quad \text{s.t.} \quad \mathbf{A}\boldsymbol{x} = \lambda \boldsymbol{x} \qquad \mathbf{A} = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & \frac{1}{2} \\ 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & \frac{1}{2} & 0 & 0 \end{pmatrix} \quad \text{with } \lambda\text{'s} = 1$$

Webs A and B are **disconnected graphs** with **non-unique rankings**

Issues with simple model - disconnected graphs (2/2)



We let

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix} \quad \text{s.t. } \mathbf{A}\boldsymbol{x} = \lambda\boldsymbol{x} \qquad \mathbf{A} = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & \frac{1}{2} \\ 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & \frac{1}{2} & 0 & 0 \end{pmatrix} \quad \text{with } \lambda\text{'s} = 1$$

Webs A and B are **disconnected graphs** with **non-unique rankings**

By setting $\mathbf{A}\boldsymbol{x} = \lambda\boldsymbol{x}$ or $(\mathbf{A} - \lambda\, \mathbf{I}\,)\boldsymbol{x} = \mathbf{0}$, we find there are **2** eigenvectors

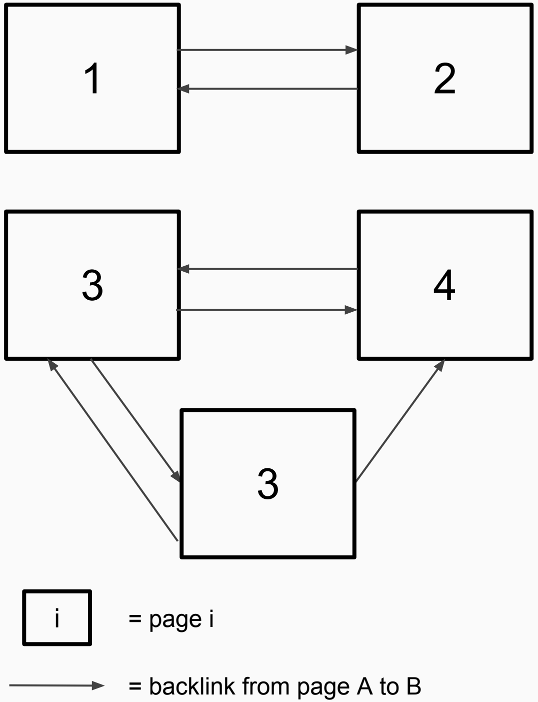
$$\boldsymbol{x_1} = \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} \quad \& \quad \boldsymbol{x_2} = \begin{pmatrix} 0 \\ 0 \\ 2 \\ 1.5 \\ 1 \end{pmatrix}$$

Non-unique rankings

- $\mathbf{dim}[\mathbf{V}_1(\mathbf{A})] > r$, where r = # of sub-webs

Source: Bryan & Leise, 2006

PageRank modification (1/2)



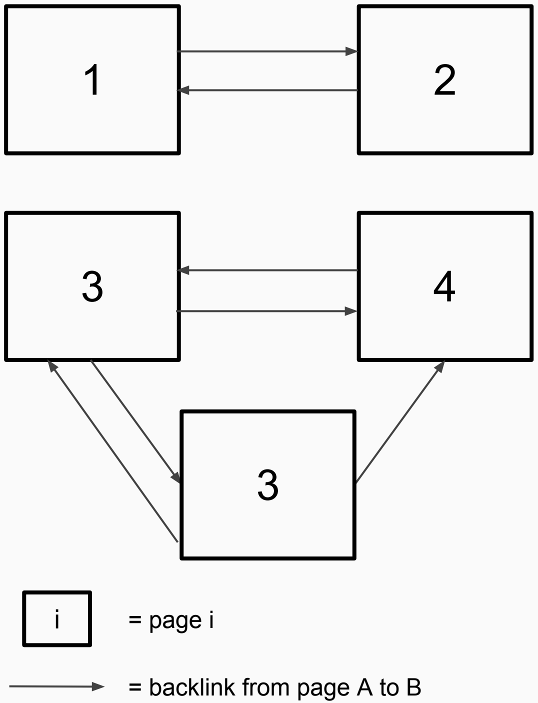
Consider model as “random surfer” instead of “random walk”

- Surfer will often follow links from one page to another (represented by matrix A from before)
- A small *m*% of the time, surfer will choose an arbitrary page
  - *m* called the **damping factor**

New transition matrix is still positive, column-stochastic

Source: Bryan & Leise, 2006

PageRank modification (2/2)



Consider model as “random surfer” instead of “random walk”

- Surfer will often follow links from one page to another (represented by matrix A from before)
- A small m% of the time, surfer will choose an arbitrary page
  - m called the **damping factor**

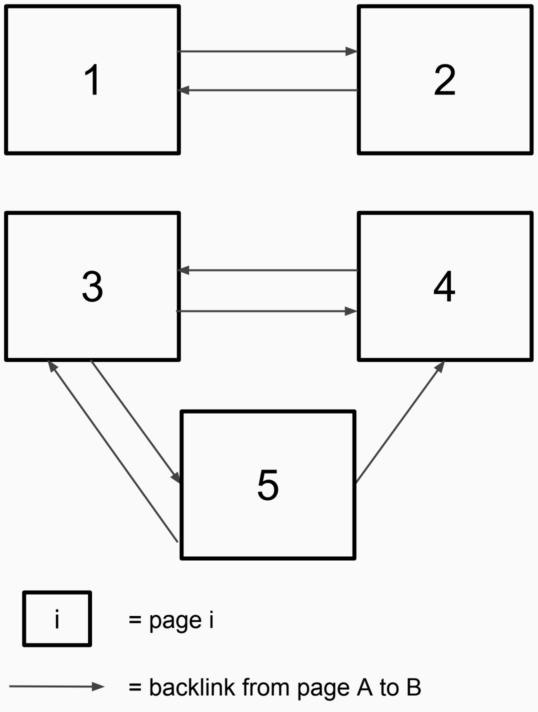
New transition matrix is still positive, column-stochastic

$\mathbf{M} = (1 - m)\mathbf{A} + m\mathbf{S}$

- $m \in [0,1]$ (Google reportedly uses $m = 0.15$)
  - m is called the **damping factor**
- $\mathbf{S}$ = $n \times n$ matrix with all entries $1/n$

Source: Bryan & Leise, 2006

Comparing PageRank scores from matrix A and matrix M (5x5)



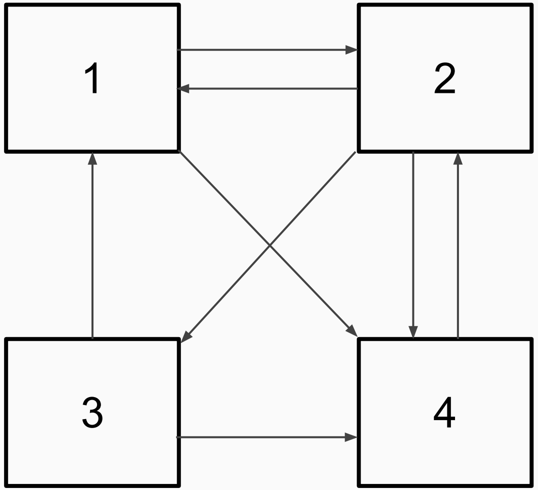
Matrix A:

$$\begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & \frac{1}{2} \\ 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & \frac{1}{2} & 0 & 0 \end{pmatrix}$$

Matrix M:

$$\begin{pmatrix} 0.03 & 0.88 & 0.03 & 0.03 & 0.03 \\ 0.88 & 0.03 & 0.03 & 0.03 & 0.03 \\ 0.03 & 0.03 & 0.03 & 0.88 & 0.455 \\ 0.03 & 0.03 & 0.455 & 0.03 & 0.455 \\ 0.03 & 0.03 & 0.455 & 0.03 & 0.03 \end{pmatrix} \quad x = \begin{pmatrix} 0.2 \\ 0.2 \\ 0.26 \\ 0.2 \\ 0.14 \end{pmatrix}$$

Comparing PageRank scores from matrix A and matrix M (4x4)



Matrix A:

$$\begin{bmatrix} 0 & \frac{1}{3} & \frac{1}{2} & 0 \\ \frac{1}{2} & 0 & 0 & 1 \\ 0 & \frac{1}{3} & 0 & 0 \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{2} & 0 \end{bmatrix} \qquad \boldsymbol{x} = \begin{bmatrix} 0.194 \\ 0.387 \\ 0.129 \\ 0.290 \end{bmatrix}$$

Matrix M:

$$\begin{bmatrix} 0.037 & 0.321 & 0.463 & 0.037 \\ 0.463 & 0.037 & 0.037 & 0.889 \\ 0.037 & 0.321 & 0.037 & 0.037 \\ 0.463 & 0.321 & 0.463 & 0.037 \end{bmatrix} \qquad \boldsymbol{x} = \begin{bmatrix} 0.202 \\ 0.368 \\ 0.142 \\ 0.288 \end{bmatrix}$$

i = page i

→ = backlink from page A to B

Note the rankings are the same in a connected web

Is the modification appropriate?

Must test for **column-stochasticity**, **existence** and **uniqueness**

- Is M column-stochastic?
- Existence: does every column-stochastic matrix have 1 as an eigenvalue?
- Uniqueness: is $V_1(\mathbf{M})$ one-dimensional?

Perron-Frobenius Theorem: a real square matrix has a unique largest eigenvalue, with the corresponding eigenvector having only positive or negative entries

Source: Wikipedia

Computation

Numerical analysis of PageRank (1/2)

In real life, matrix M may be a billion-dimensional matrix

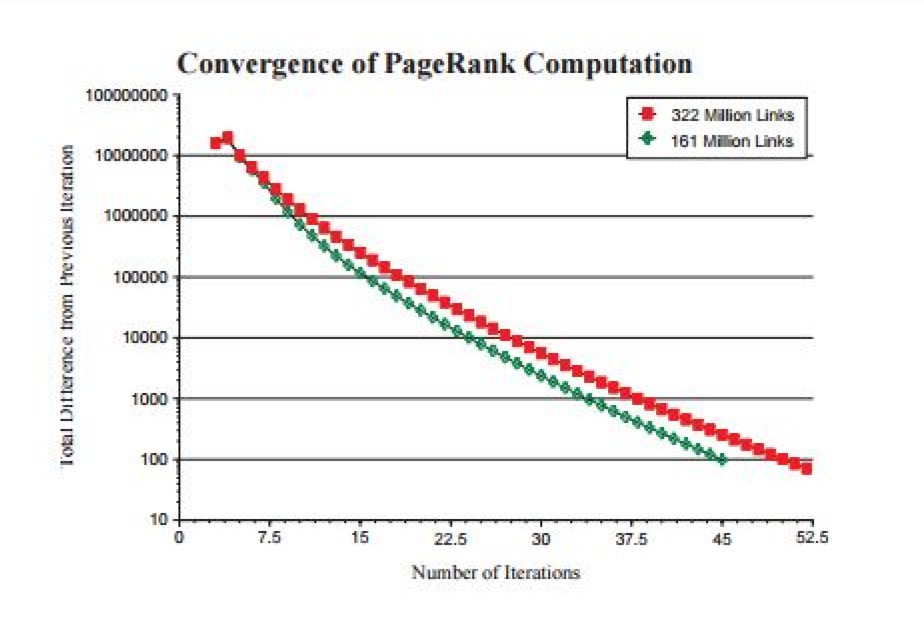
- Computing an eigenvector requires iteration

Power Method Convergence Theorem

- Let x_0 represent the initial vector. The sequence x_0, $\mathbf{M}x_0$, $\mathbf{M}^2x_0$, …, $\mathbf{M}^kx_0$ converges to the eigenvector $\boldsymbol{x}$

Source: Tanase and Radu, 2009

Numerical analysis of PageRank (2/2)



Source: Brin & Page, 1998

Python functions

```
def getM(A,m,n):
    S = ones((n,n))/n
    return (1-m)*A + m*S

def powerMethod(A,x0,k):
    for i in range(0,k):
        x0 = dot(A,x0)
    return x0
```

getM: computes M matrix

powerMethod: given initial vector x_0, number of iterations k, and matrix **A**, $\mathbf{A}^k x_0$ is calculated

Code adapted from http://glowingpython.blogspot.com/2011/05/four-ways-to-compute-google-pagerank.html

Implications and future

How to get a high PageRank?

Control a site that has many affiliates referring back it

E.g. NYT's home page

Create something that other sites will use, and have a link back

E.g. "Powered by …"

Beware of risky methods

Ex: RapGenius, December 2013

Source: Rogers

Implications

Google created to test PageRank

- “25 billion dollar eigenvector”
- 64% of market share, valued at ~$500B (as of July 2015)

Outside applications

- Ecological modeling
- Network modeling
- Recommendation systems

Currently, PageRank is one of many algorithms (N=200+) used by Google to filter search results

Sources: Comscore, 2015

About us

Varun Kumar

Marco Groenendaal

Source: LinkedIn

Appendix

Sources

K. Bryan and T. Leise, *The $25,000,000,000 Eigenvector: The Linear Algebra Behind Google,* https://www.rose-hulman.edu/~bryan/googleFinalVersionFixed.pdf, (accessed September 30, 2015).

S. Brin and L. Page, *The Anatomy of a Large Scale Hypertextual Web Search Engine*, http://infolab.stanford.edu/~backrub/google.html, (accessed September 28, 2015).

S. Brin and L. Page, *The PageRank Citation Ranking: Bringing Order to the Web*, http://ilpubs.stanford.edu:8090/422/1/1999-66.pdf, (accessed September 28, 2015).

Aerostudents, *Linear Algebra Summary*, http://aerostudents.com/files/linearAlgebra/linearAlgebraFullVersion.pdf, (accessed September 30, 2015).

From Wolfram Alpha:

Weisstein, Eric W. "Markov Chain." From *MathWorld*--A Wolfram Web Resource. http://mathworld.wolfram.com/ MarkovChain.html

Weisstein, Eric W. "Random Walk." From MathWorld--A Wolfram Web Resource. http://mathworld.wolfram.com/ RandomWalk.htm

Weisstein, Eric W. "Stochastic Matrix." From MathWorld--A Wolfram Web Resource. http://mathworld.wolfram.com/StochasticMatrix.htmll

Sources

I. Rogers, *The Google PageRank Algorithm and How It Works,* http://www.cs.princeton.edu/~chazelle/courses/BIB/pagerank.htm, (accessed September 30, 2015).

Comscore, *comScore Releases June 2015 U.S. Desktop Search Engine Rankings,* http://www.comscore.com/Insights/Market-Rankings/comScore-Releases-June-2015-US-Desktop-Search-Engine-Rankings, (accessed October 3, 2015).

R. Tanase and R. Radu, *Lecture #3: PageRank Algorithm - The Mathematics of Google Search*. http://www.math.cornell.edu/~mec/Winter2009/RalucaRemus/Lecture3/lecture3.html, (accessed October 2, 2015).

A little bit of humor

Web Page	PageRank (average is 1.0)
Download Netscape Software	11589.00
http://www.w3.org/	10717.70
Welcome to Netscape	8673.51
Point: It's What You're Searching For	7930.92
Web-Counter Home Page	7254.97
The Blue Ribbon Campaign for Online Free Speech	7010.39
CERN Welcome	6562.49
Yahoo!	6561.80
Welcome to Netscape	6203.47
Wusage 4.1: A Usage Statistics System For Web Servers	5963.27
The World Wide Web Consortium (W3C)	5672.21
Lycos, Inc. Home Page	4683.31
Starting Point	4501.98
Welcome to Magellan!	3866.82
Oracle Corporation	3587.63

Table 1: Top 15 Page Ranks: July 1996

Source: Brin and Page, 1998