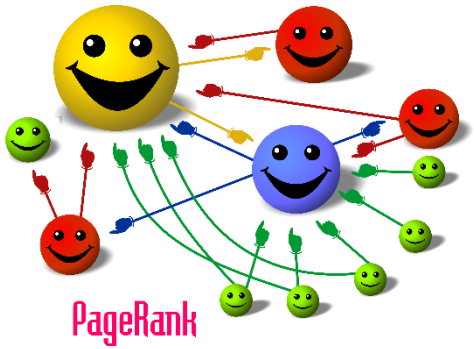
# PageRank

Reading:

Supplementary Notes on PageRank



# Importance of Ranking Web Pages

Ranking of web pages is very important for users of the Web as well as business.

- For a given user query, there could be thousands, millions and even billions of web pages that satisfy the query. For example, querying “algorithms” yields about 415,000,000 results. Querying “Computer Science” yields about 2,920,000,000 results
- The ranking of the web pages effects the order in which their hyperlinks are listed.
- Most people only look at the first few pages of results for their query, sometimes only looking at the hyperlinks listed at the top of the first page.
- From the user’s point of view, it is important that web pages of “high quality”, “high prestige” get listed first.
- For a business, it makes a great deal of difference whether a hyperlink to their web page shows up near the beginning of the list, for example, near the top of the first page, potentially bringing in lots of new business.

# PageRank


*PageRank*, was developed in 1996 at Stanford by Larry Page and Sergey Brin, the cofounders of Google.

# Ph.D. Students at Stanford

Page and  Brin were both Ph.D. students at Stanford, when they came up with the killer application of PageRank and pioneered the Google Search Engine.

# Idea  behind PageRank

*PageRank* assigns a measure of "prestige" or ranking (*PageRank*) to each web page, which is independent of any query. It is defined using a digraph **based on the hyperlink structure of the web** called the *web digraph*.

# Business Flop?

They tried to sell their search engine idea using PageRank to YAHOO!

But it was rejected

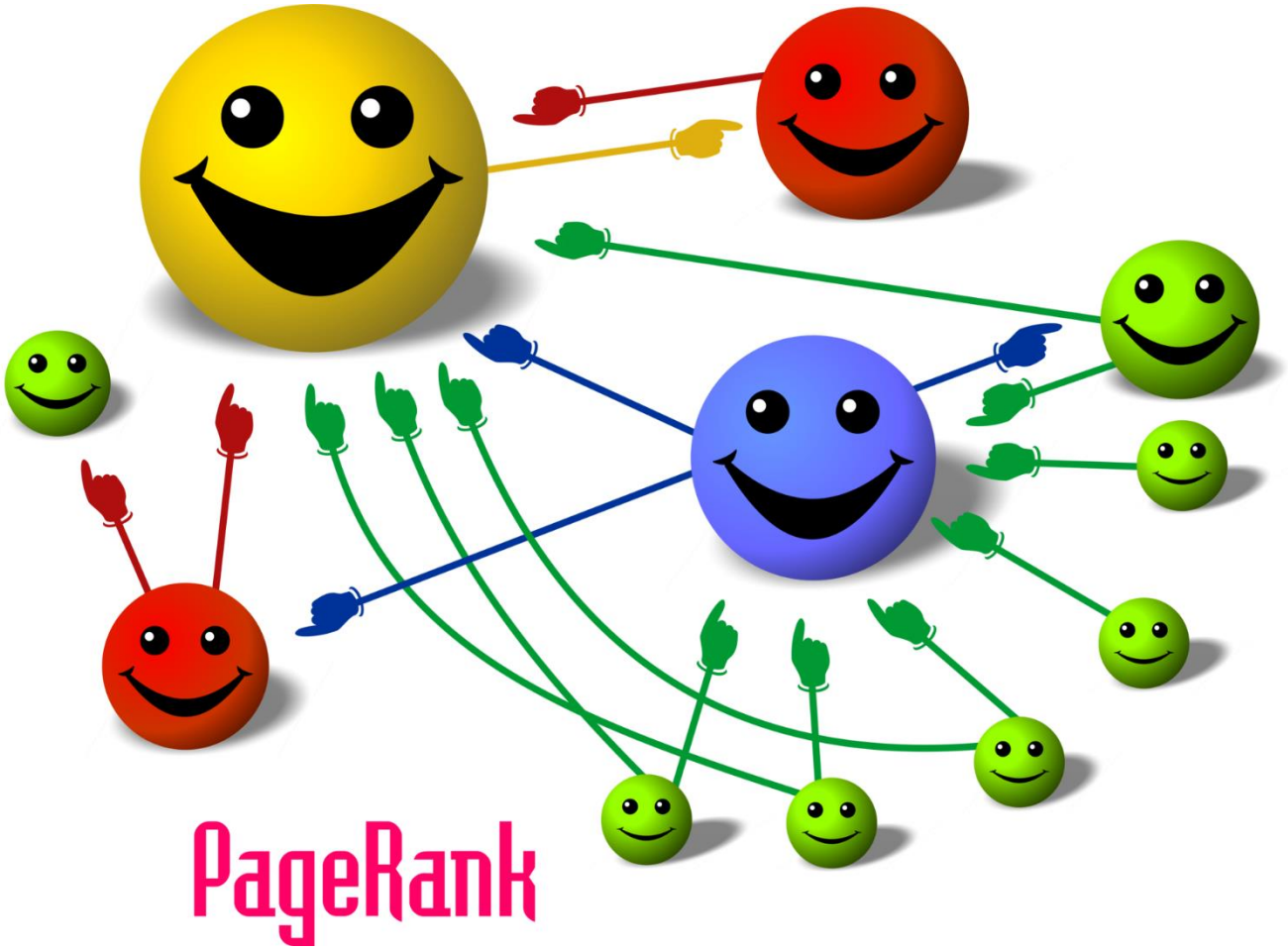
# Founding of Google

- After the rejection by Yahoo, Page and Brin started their own company, which they called Google.
- One of the first investors in their company was a professor at Stanford, David Cheriton, who had received his M.S. and  Ph.D. degrees in computer science from the University of Waterloo.
- Cheriton later donated $25 million to support graduate studies and research in the School of Computer Science, subsequently renamed David R. Cheriton School of Computer Science, at the University of Waterloo

# Web Digraph

- The **web digraph** $W$
- Vertex set $V(W)$ consists of **web pages**
- Edge set $E(W)$ corresponds to **hyperlinks**, that is, an edge is included from page $p$ to page $q$ whenever there is a hyperlink reference (href) in page $p$ to page $q$.

# In its simplified form PageRank of a web page is measured by its in-degree in *W*



# Drawback with Simplistic Definition

Using just the in-degree of a web page *p* as its rank has two weakness:

- Web pages *q* that contain a hyperlink to *p* may have different measures of prestige.
- Web pages *q* that contain a hyperlink to *p* may have different out-degrees.

If *q* has lower prestige, we don't want to count it as heavily.

Similarly, we don't want to count it as much if it has high out-degree, i.e., it includes a lot of hyperlink references.

# Formula for PageRank

The PageRank $R[p]$ of web page $p$ satisfies:

$$R[p] = \sum_{q \in N_{in}(p)} \frac{R[q]}{d_{out}(q)} .$$

where $d_{out}(q)$ is the **out-degree** of page $q$, or equivalently the **number $h(q)$ of hyperlink references** that $q$ contains.

# PageRank has Myriad Applications

PageRank has been used for many other applications besides Google including applications to

- Bibliometrics
- Social and information network analysis
- Link prediction and recommendation.
- Systems analysis of road networks
- Gene Searching in Biology: an app called ToppGene Suite was developed at UC and Cincinnati Children's Hospital by Anil Jegga https://toppgene.cchmc.org/
- Index called pagerank-index (Pi) for quantifying the scientific impact of researchers

# PageRank Concept Existed Before


The concept behind PageRank was not invented by Page and Brin, but was first applied by them to the hyperlink structure of the Web in the design Google. In fact, in their famous paper

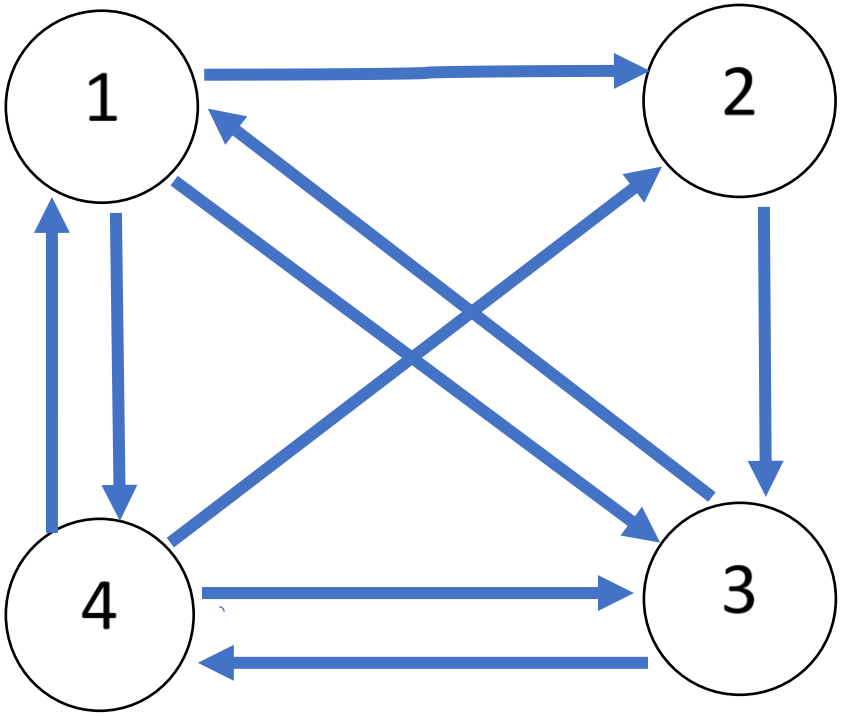
The Anatomy of a Large-Scale Hypertextual Web Search Engine

they only devote one paragraph to PageRank. To quote their paper:

“Academic citation literature has been applied to the web, largely by counting citations or backlinks to a given page. This gives some approximation of a page's importance or quality.”

# Computing PageRank for Mini Web Digraph $W = (E,V)$

=



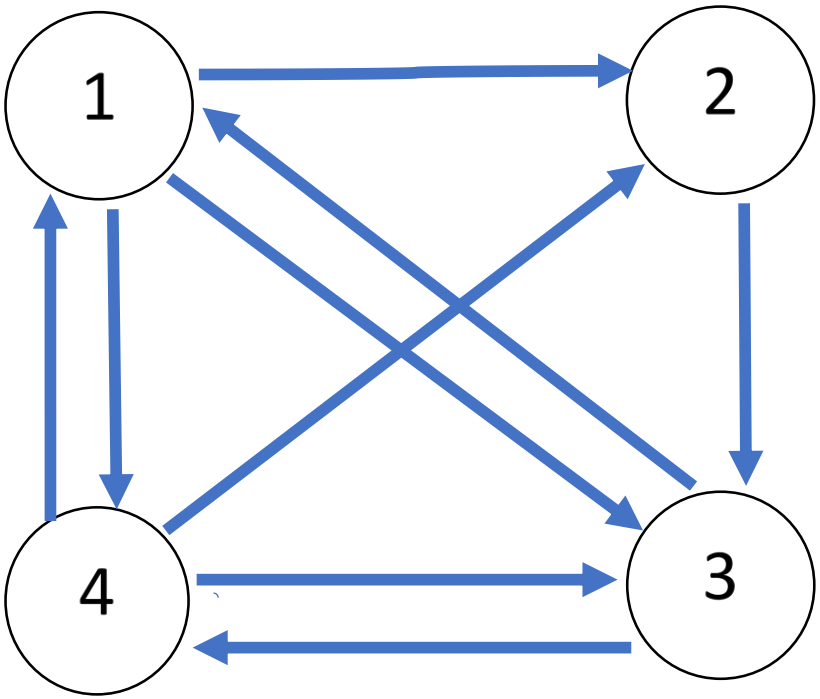
# Page Rank Equations:

$$R_1 = \frac{1}{2}R_3 + \frac{1}{3}R_4$$

$$R_2 = \frac{1}{3}R_1 + \frac{1}{3}R_4$$

$$R_3 = \frac{1}{3}R_1 + R_2 + \frac{1}{3}R_4$$

$$R_4 = \frac{1}{3}R_1 + \frac{1}{2}R_3$$

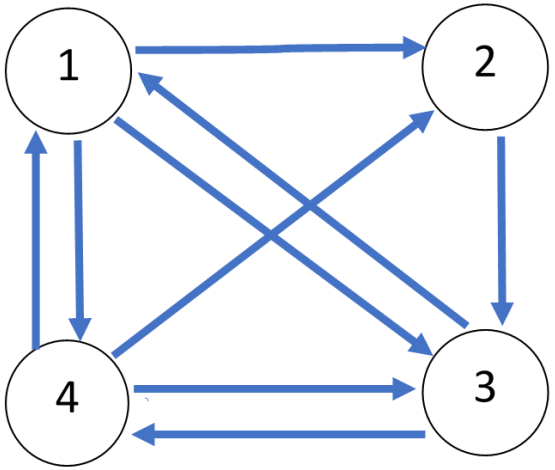


# Expressing Linear Equations in Matrix Form

$$R = \begin{pmatrix} 0 & 0 & \frac{1}{2} & \frac{1}{3} \\ \frac{1}{3} & 0 & 0 & \frac{1}{3} \\ \frac{1}{3} & 1 & 0 & \frac{1}{3} \\ \frac{1}{3} & 0 & \frac{1}{2} & 0 \end{pmatrix} R, \qquad \text{where } R = \begin{bmatrix} R_1 \\ R_2 \\ R_3 \\ R_4 \end{bmatrix}$$

or equivalently,

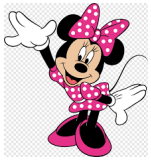
$$R = \begin{pmatrix} 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ 0 & 0 & 1 & 0 \\ \frac{1}{2} & 0 & 0 & \frac{1}{2} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 \end{pmatrix}^T R$$



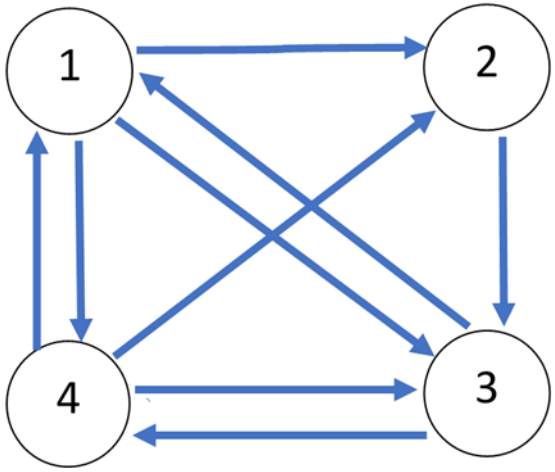
# Random Walk Matrix

Let *B* be the matrix for a random walk on *W*, i.e.,

$$B[p,q] = \begin{cases} \frac{1}{d_{out}(p)} , & pq \in E(W), \\ 0 , & \text{otherwise.} \end{cases}$$

For  Web, *B* is the matrix for a **random walk**!

$$\begin{pmatrix} 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ 0 & 0 & 1 & 0 \\ \frac{1}{2} & 0 & 0 & \frac{1}{2} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 \end{pmatrix}$$

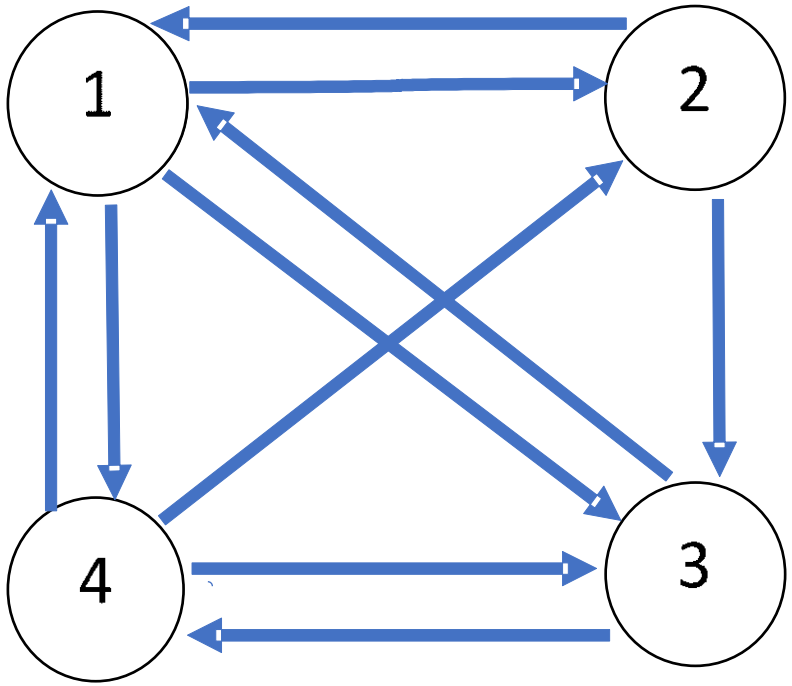


PSN. For the mini-web below

a) Give the Page Rank equations

b) Give the associated matrix equation

c) Give transpose matrix and verify it is the matrix for a random walk.



# PageRank is Principal Eigenvector of Matrix of a Random Walk on *W*

Letting be the PageRank *R* be the vector or array, i.e., *R*[*p*] is the PageRank of Page p, the formula for PageRank is

$$R = B^{T}R.$$

*R* is an eigenvector of the matrix *B* (and $B^{T}$) for the eigenvalue 1. Since 1 is the largest eigenvalue of B, *R* is **principal eigenvector**.

# Standard Linear Algebra Algorithms to Compute Principal Eigenvector

Using standard linear algebra algorithms to compute Principal Eigenvector is **prohibitive** for large matrices.

At the present time, the last realistic estimate of the World Wide Web was about **19.2 billion** web pages. It is not possible to store a matrix of this size, because the number of entries is **astronomical**!

368,640,000,000,000,000,000

# Efficient Algorithm by Iterating

$$R_i = B^T R_{i-1},\ i = 1, 2, \ldots$$

or equivalently,

$$R_i[p] = \sum_{q \in N_{in}(p)} \frac{R_{i-1}[q]}{d_{out}(q)},\ i = 1, 2, \ldots$$

where $R_0$ can chosen to be a vector of nonnegative reals whose entries sum to 1.

# Random Walk Interpretation of PageRank

- By a simple induction argument, we obtain

$$R_i = (B^{\mathsf{T}})^i R_0,\ i = 1, 2, \ldots$$

- By a theorem of Markov, $B^i[p,q]$ equals the probability of a random walk on $W$ starting a page $p$ and ending up at page $q$ after $i$ steps. Thus,
- $(B^{\mathsf{T}})^i[p,q] = B^i[q,p]$ = probability that a random walk starting at page $q$ will end up at page $p$ after $i$ steps.

# Equivalent Interpretation

$B^i[p][q]$ is the probability that an aimless web surfer starting page $p$ reaches $q$ in $i$ steps (by following a path of $i$ hyperlinks).

Equivalently, $(B^T)^i[p][q]$ is the probability that the aimless web surfer starting a page $q$ will reach $p$ after $i$ steps.

# Random (Aimless) Web Surfer and PageRank

Since the entries of $R_0$ are positive and sum to 1, $R_0$ determines a probability distribution on the set of pages $V(W)$, where $R_0[q]$ is the probability that the aimless surfer begins surfing from page $q$.

Then $R_i[p] = (B^T)^i R_0[p]$ is the probability the surfer will end up on page $p$ after $i$ steps.

# Interesting Special Case

For a particular page $q$, e.g., your web page, set $R_0[q] = 1$ and set $R_0[p] = 0$, for all $p \neq q$.

Then $R_i[p] = (B^T)^i R_0[p]$ is the probability that a random surfer starting at page $q$ will end up at page $p$ after $i$ steps.

# Conditions for convergence

The vector $R_i$ will not necessarily converge to the principle eigenvector $R$ unless the digraph $W$ satisfies certain conditions.

**Condition 1. *W*** is **strongly-connected.** There is a directed path from $p$ to $q$ for every two vertices $p$ and $q$.

**Condition 2. *W* is aperiodic**. There is some integer $N$, such that for all $k \geq N$, $W$ contains a closed walk of length $k$ starting at any given vertex.

# Damping Factor

- The actual World Wide Web Digraph is neither strongly-connected nor aperiodic.
- To ensure that the iteration for PageRank converges it is necessary to introduce a **damping factor**.
- Let *n* denote the number of nodes of the web digraph W. The PageRank *R*[*p*] of a web page *p* is given by:

$$R[p] = \frac{1 - d}{n} + d \sum_{q \in N_{in}(p)} \frac{R[q]}{d_{out}(q)}$$

where *d* is the damping factor between 0 and 1.

# Iteration Converges with Damping Factor

- The damping factor is equivalent to adding edges for every pair of web pages $(p, q)$ giving it a the very small probability $\frac{1-d}{n}$ and slightly reducing the probability of each original edge having tail $q$ from the $\frac{1}{d_{out}(q)}$ to $\frac{d}{d_{out}(q)}$
- The underlying digraph is now complete, so that it is necessarily strongly connected and aperiodic.
- This means the iteration for PageRank will always converge when a damping factor is added!

# Computing PageRank in Practice for the World Wide Web

- In practice compute PageRank by iterating

$$R_i[p] = \frac{1-d}{n} + d \sum_{q \in N_{in}(p)} \frac{R_{i-1}[q]}{d_{out}(q)} \, , i = 1, 2, \ldots$$

- Empirical experiments have shown that acceptable ranking functions $R_i$ are achieved in 52 iterations for about 322 million hyperlinks.

- Taking the damping factor *d* to be between .8 and .9 has been found to work well in practice.

# Random Web Surfer with Damping Factor

- As before the value of PageRank after *i* iterations $R_i[p] = (B^T)^i R_0[p]$ is the probability that a random surfer will end up on page *p* after *i* steps.
- The difference is that the surfer can randomly jump from a page *q* to an arbitrary page, but with a very small probability, i.e., $\frac{1-d}{n}$.
- Otherwise, the surfer randomly goes with equal probability ( $\frac{d}{d_{out}(q)}$ ) to a page in the out-neighborhood of *q, i.e.,* clicks on a hyperlink on page *q*.

What do frogs say that surf the internet?

Reddit reddit.