Lecture 18: Numerical Linear Algebra – Page Rank

Leili Rafiee Sevyeri

Based on lecture notes by me and many previous CS370 instructors

Winter 2024

Cheriton School of Computer Science
University of Waterloo

Page Rank: The Not-So secret sauce behind Google.

A Bit of History – Ranking Strategy

Websites mostly ranked by counting keywords on each site, with some variations.

- Clearly easy to "cheat".
- Often gave poor results.

Alternative: Yahoo was a big human-curated directory structure.

No consistently dominant search engine.

A Bit of History – The Dawn of Google

Around 1998, along came Stanford PhD students, Sergey Brin and Larry Page.

Rough idea: If many web pages link to your website, there must be a consensus that it is important.

A simple analogy...

Good indicator:

- Everybody else tells you *Joe's Used Cars* is trustworthy.

Poor indicators:

- Joe himself constantly tells you he's honest.
- Joe publishes ads saying he's really reliable.

Analogy #2: Paper Citations

Academic success is sometimes measured similarly.

- I write a research paper, citing earlier work.
- If my paper then gets cited **many times** by other people's papers, this suggests my paper was influential.

This provided some inspiration.

e.g. The original paper describing PageRank now has 11832 citations (Google Scholar).

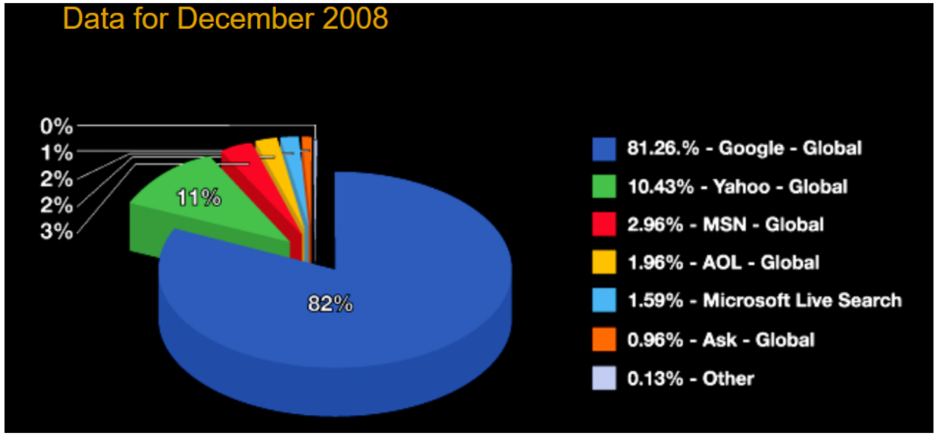
The **PageRank citation ranking**: **Bringing order** to the **web**.
L Page, S Brin, R Motwani, T Winograd - 1999 - ilpubs.stanford.edu
… 1.2 **PageRank** In **order** to measure the relative importance of **web** pages, we propose **PageRank**, a method for computing a **ranking** for every **web** page based on the graph of the …
☆ Save 99 Cite Cited by 17037 Related articles All 16 versions »

Google Homepage in September 1998

Page and Brin launched Google to commercialize the PageRank algorithm.

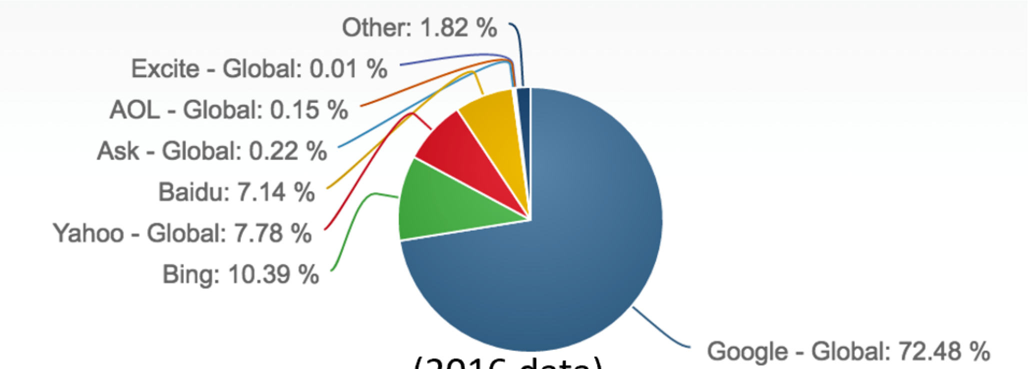
Google rapidly took over the search engine market due to its superior results.

That was the end of the story until 2009...



And then, along came Bing.

(And nothing much changed.)



(2016 data)

Key Takeaway

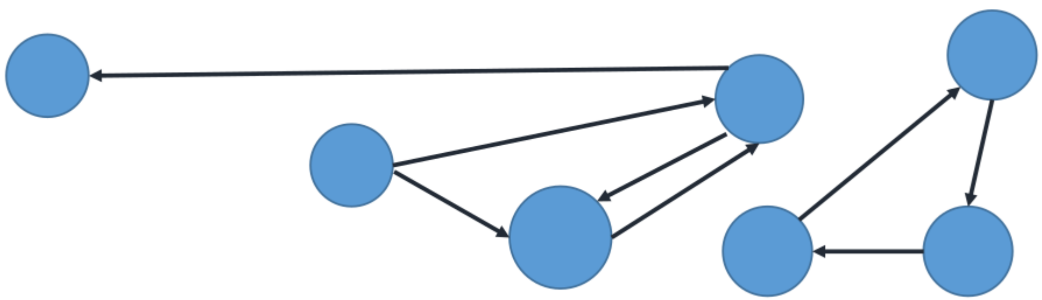
Numerical linear algebra could earn you \$35 billion by age 43 (if you don't mind posing for dorky photos).



PageRank Idea: From Links to Importance

Use a directed graph to model the problem

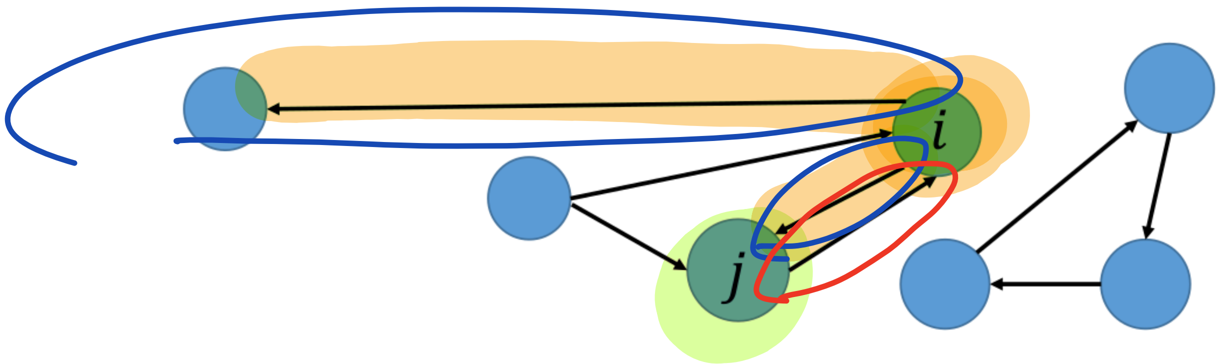
Clearly the link structure *between* pages provides some useful indicator.



Page and Brin turned this vague idea into a concrete ***importance metric***, using tools of numerical linear algebra.

Web Links as a Graph

We represent the web's structure as a **directed graph.**



Nodes (circles) represent pages.

Arcs (arrows) represent links from one page to another.

We will use **degree** to refer to a node's outdegree, the number of arcs leaving that node.

e.g. $\deg(j) = 1$, $\deg(i) = 2$.

Adjacency Matrix

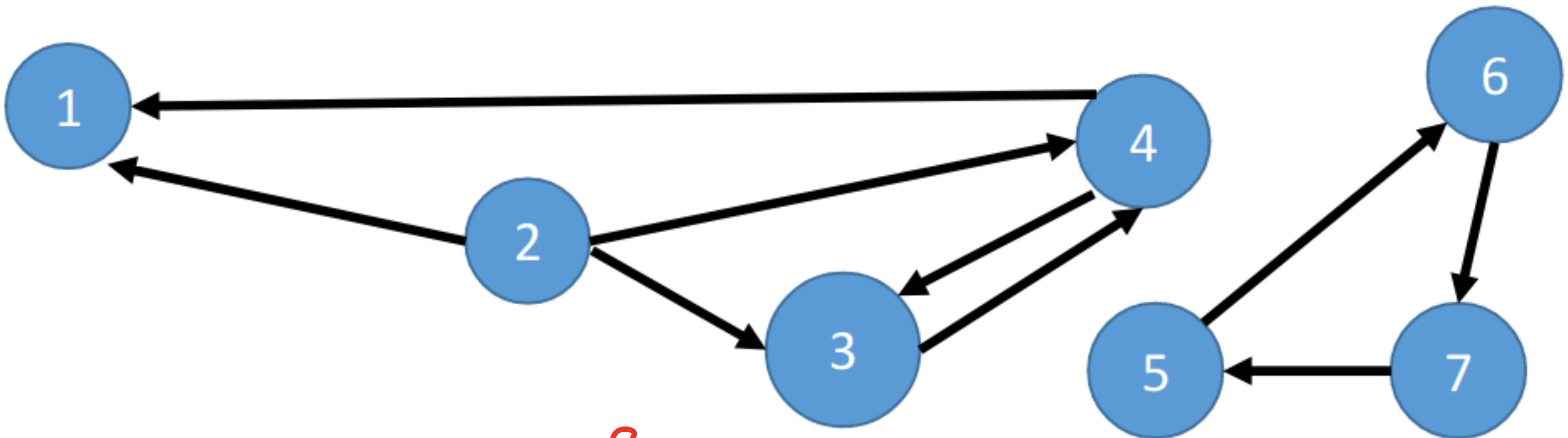
To store our directed graph, we can use a kind of adjacency matrix, G.

$$G_{ij} = \begin{cases} 1, & \text{if link j} \to \text{i exists} \\ 0, & \text{otherwise} \end{cases}$$

Then the (out)degree for node q is the sum of entries in column q.

Notice: Matrix G is not necessarily symmetric about the diagonal!

Example Adjacency Matrix



from

Question: What is the adjacency matrix of this graph?

$$G = \begin{array}{c c} \begin{array}{ccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 \end{array} & \\ \begin{bmatrix} 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix} & \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \end{array} \end{array}$$

to

Interpreting Links as Votes

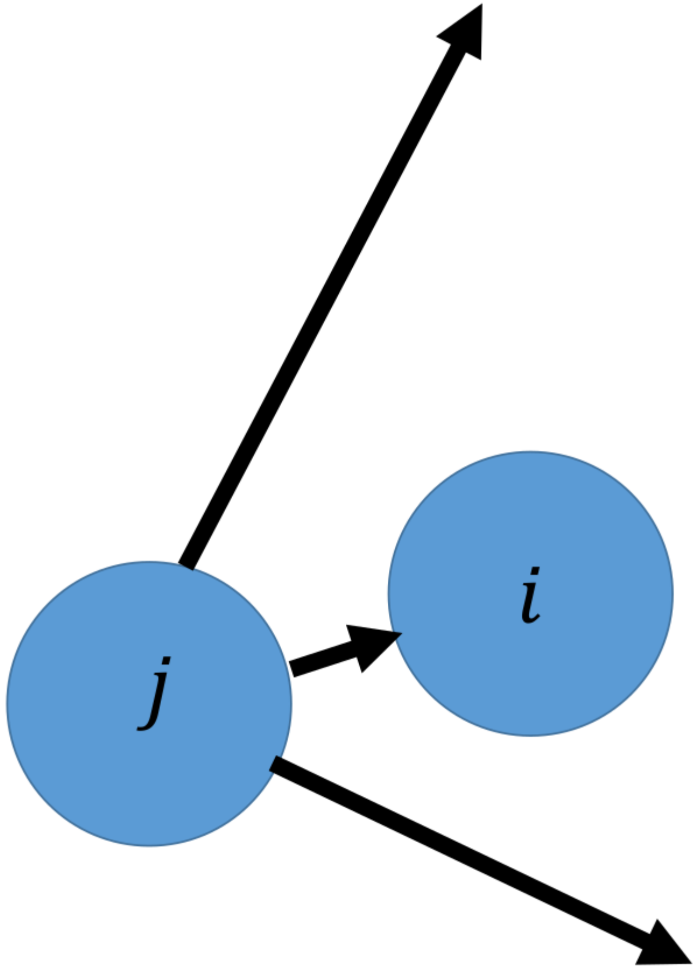
If page j links to page i, this is considered a "vote" by j that i is "important".

Outgoing links of a page j have equal influence, so the importance that j "gives" to i is:

$$\frac{1}{\deg(j)}$$

deg(j) = 3

e.g., in the diagram, j gives a 1/3 vote to i.



Global Importance

So: If page i has many incoming links, it is probably important.

What if page i has just one incoming link, but the link is from page j, and j **has many incoming links?**
Then i is probably fairly important too!

The Random Surfer Model

Imagine an internet user who starts at a page, and **follows links at random** from page to page for K steps.

They will "probably" end up on important pages more often!

Then, select a new start page, and follow K random links again. Repeat the process R times, starting from each page.

At the end, we estimate overall importance as:

$$\text{Rank(page i)} = \big(\text{Visits to page i}\big) / \underbrace{\big(\text{Total visits to all pages}\big)}_{R \times K}$$

Random Surfer Algorithm

$Rank(m) = 0, m = 1, \ldots, R$
For $m = 1, \ldots, R$
    $j = m$
    For $k = 1, \ldots, K$
        $Rank(j) = Rank(j) + 1$
        Randomly select outlink l of page j
        $j = l$
    EndFor
EndFor
$Rank(m) = Rank(m)/(K \times R),\ m = 1, \ldots R$

Random Surfer Criticisms

Potential issues with this algorithm?

- The number of real web pages is monstrously huge: 1 billion-ish unique hostnames; **many** iterations (large K, R) needed.
- Number of steps taken per random surf sequence must be large, to get a representative sample.
- What about dead end links? (Stuck on **one** page!)
- What about cycles in the graph? (Stuck on a **closed subset** of pages!)

Clearly, better strategies are needed.

A Markov Chain Matrix

Let's think in terms of probabilities instead.

Let P be a (large!) matrix of probabilities, where P_{ij} is the probability of randomly transitioning from page j to page i.

$$P_{ij} = \begin{cases} \dfrac{1}{\deg(j)}, & \text{if link j} \rightarrow \text{i exists} \\ 0, & \text{otherwise} \end{cases}$$

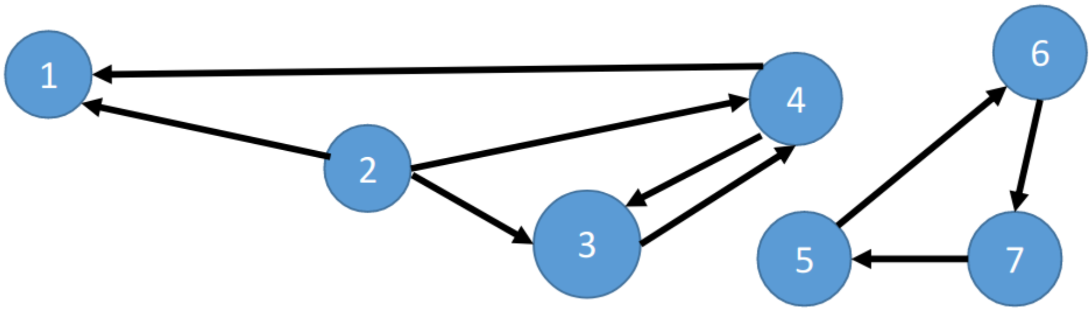
Markov Chain Matrix

We can build this matrix P from our adjacency matrix G.

Divide all entries of each column of G by the column sum (out-degree of the node).

From

$$\text{To}\quad \begin{array}{cc} & \begin{array}{ccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 \end{array} \\ \begin{array}{c} 1\\ 2\\ 3\\ 4\\ 5\\ 6\\ 7 \end{array} & \begin{bmatrix} 0 & 1/3 & 0 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1/3 & 0 & 1/2 & 0 & 0 & 0 \\ 0 & 1/3 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix} \end{array}$$



$$
\begin{array}{cccccccccc}
 & & & & \text{from} & & & & & \\
 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & & \\
\end{array}
$$

$$
G = \left[\begin{array}{ccccccc}
0 & 1 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 1 & 0 & 0 & 0 \\
0 & 1 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 1
\end{array}\right]
\begin{array}{c}
1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7
\end{array}
\ \text{to}
$$

$$
\begin{array}{cccccccc}
\text{add} & \downarrow & +\downarrow & +\downarrow & +\downarrow & +\downarrow & +\downarrow & +\downarrow \\
 & 0 & 3 & 1 & 2 & 1 & 1 & 1
\end{array}
$$

$$
P = \left[\begin{array}{ccccccc}
0 & 1/3 & 0 & 1/2 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1/3 & 0 & 1/2 & 0 & 0 & 0 \\
0 & 1/3 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 0
\end{array}\right]
$$

Dead Ends

To deal with dead-end links, we will simply "teleport" to a new page at random!

$$\begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \to \begin{bmatrix} 1/R \\ 1/R \\ \vdots \\ 1/R \end{bmatrix}$$

Mathematically, we define a column vector d such that:

d^T as a row vector

$$d_i = \begin{cases} 1, & \text{if } \deg(i) = 0 \\ 0, & \text{otherwise} \end{cases}$$

→ Column of zeros

and vector $e = [1, 1, 1, \ldots, 1, 1]^T$ be a column vector of ones.

Then if R is the number of pages, we augment P to get P' defined by:

$$P' = P + \frac{1}{R} e d^T$$

Can you see why this gives the desired effect?

Assume G is $\begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$

So, P is $\begin{bmatrix} 0 & 1/3 & 0 & 0 \\ 0 & 1/3 & 0 & 1/2 \\ 0 & 1/3 & 0 & 0 \\ 0 & 0 & 0 & 1/2 \end{bmatrix}$

Column 1 = Zero

Column 3 = Zero

So, $d = \begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 1 & 0 \end{bmatrix}^T$

and $e = \begin{bmatrix} 1 & 1 & 1 & 1 \end{bmatrix}^T$, and $\frac{1}{R} = \frac{1}{4}$

So, $\frac{1}{R} e d^T = \frac{1}{R} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 1 & 0 \end{bmatrix}$

$$= \begin{bmatrix} 1/R & 0 & 1/R & 0 \\ 1/R & 0 & 1/R & 0 \\ 1/R & 0 & 1/R & 0 \\ 1/R & 0 & 1/R & 0 \end{bmatrix}$$

Column 1 Column 3

$$P' = P + \frac{1}{R} e d^{T} = \begin{bmatrix} 1/4 & 1/3 & 1/4 & 0 \\ 1/4 & 1/3 & 1/4 & 1/2 \\ 1/4 & 1/3 & 1/4 & 0 \\ 1/4 & 0 & 1/4 & 1/2 \end{bmatrix}$$

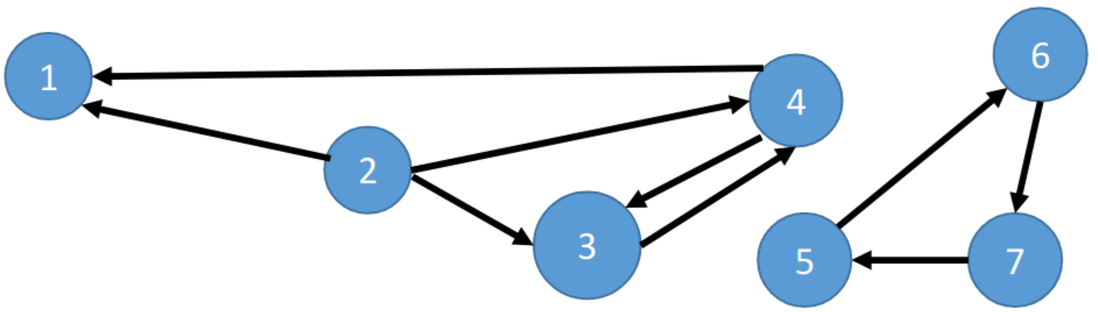
So, we have no dead ends.

Dead Ends

The **matrix** $\frac{1}{R}ed^T$ is a matrix of probabilities such that **from** any dead end page ($d_i = 1$), we transition **to** every other page with equal probability.

$$\begin{array}{ccc} & & \text{From} \\ & & \begin{array}{ccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 \end{array} \\ \text{To} & \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \end{array} & \begin{bmatrix} 1/7 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1/7 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1/7 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1/7 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1/7 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1/7 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1/7 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \end{array}$$

$$d = [1,0,0,0,0,0,0]^T$$
$$e = [1,1,1,1,1,1,1]^T$$
$$R = 7$$



Escaping Cycles

How can we apply a similar trick to escape closed cycles of pages?

Most of the time (a fraction α), follow links randomly, via P'.

Occasionally, with some (usually small) probability, $(1-\alpha)$, teleport from **any** page to **any** other page.

closed cycle

$$M = \alpha P' + (1-\alpha)\frac{1}{R}ee^T$$

$$G = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ \\ \end{array} \begin{array}{c} \begin{array}{ccc} 1 & 2 & 3 \end{array} \\ \left[\begin{array}{ccc|c} 0 & 0 & 1 & \\ 1 & 0 & 0 & * \\ 0 & 1 & 0 & \\ \hline & 0 & & * \end{array}\right] \end{array}$$

Escaping Cycles

$$ee^T = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & \ddots & & 1 \\ \vdots & & \ddots & \vdots \\ 1 & \cdots & & 1 \end{bmatrix}$$

The $\dfrac{1}{R}ee^T$ matrix looks like:

$$\frac{1}{R}ee^T = \begin{bmatrix} \frac{1}{R} & \frac{1}{R} & \cdots & \cdots & \frac{1}{R} \\ \frac{1}{R} & \frac{1}{R} & \cdots & \cdots & \frac{1}{R} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \frac{1}{R} & \frac{1}{R} & \cdots & \cdots & \frac{1}{R} \end{bmatrix}$$

Teleport randomly from one page to another with equal probability, regardless of links.

$$M = \alpha P' + \underbrace{(1-\alpha)\frac{1}{R}ee^T}$$

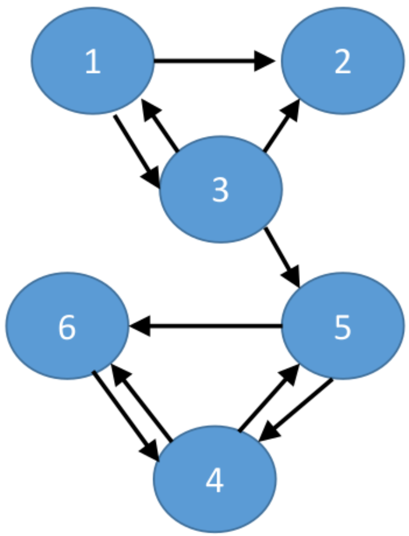
Google Matrix

We will call the combined matrix $M = \alpha P' + (1-\alpha)\frac{1}{R}ee^T$ our "Google matrix".

Most of the time this just follows links (and always teleports out of dead ends), but also occasionally teleports randomly to escape cycles. Google purportedly used $\alpha \approx 0.85$.

End of Lecture 18

Page Rank example (Notes Ex. 7.4)



$$P = \begin{bmatrix} 0 & 0 & \frac{1}{3} & 0 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{3} & 0 & 0 & 0 \\ \frac{1}{2} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{2} & 1 \\ 0 & 0 & \frac{1}{3} & \frac{1}{2} & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 \end{bmatrix}$$

$$d = [0, 1, 0, 0, 0, 0].$$

(Page 2 is a dead end!)

Random Surfing
(but gets stuck in dead ends)

Page Rank example

$$P' = P + \frac{1}{6}ed^T = \begin{bmatrix} 0 & \frac{1}{6} & \frac{1}{3} & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{6} & \frac{1}{3} & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{6} & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{6} & 0 & 0 & \frac{1}{2} & 1 \\ 0 & \frac{1}{6} & \frac{1}{3} & \frac{1}{2} & 0 & 0 \\ 0 & \frac{1}{6} & 0 & \frac{1}{2} & \frac{1}{2} & 0 \end{bmatrix}$$

$$M = \begin{bmatrix} \frac{1}{6}-\frac{1}{6}\alpha & \frac{1}{6} & \frac{1}{6}\alpha+\frac{1}{6} & \frac{1}{6}-\frac{1}{6}\alpha & \frac{1}{6}-\frac{1}{6}\alpha & \frac{1}{6}-\frac{1}{6}\alpha \\ \frac{1}{3}\alpha+\frac{1}{6} & \frac{1}{6} & \frac{1}{6}\alpha+\frac{1}{6} & \frac{1}{6}-\frac{1}{6}\alpha & \frac{1}{6}-\frac{1}{6}\alpha & \frac{1}{6}-\frac{1}{6}\alpha \\ \frac{1}{3}\alpha+\frac{1}{6} & \frac{1}{6} & \frac{1}{6}-\frac{1}{6}\alpha & \frac{1}{6}-\frac{1}{6}\alpha & \frac{1}{6}-\frac{1}{6}\alpha & \frac{1}{6}-\frac{1}{6}\alpha \\ \frac{1}{6}-\frac{1}{6}\alpha & \frac{1}{6} & \frac{1}{6}-\frac{1}{6}\alpha & \frac{1}{6}-\frac{1}{6}\alpha & \frac{1}{3}\alpha+\frac{1}{6} & \frac{5}{6}\alpha+\frac{1}{6} \\ \frac{1}{6}-\frac{1}{6}\alpha & \frac{1}{6} & \frac{1}{6}\alpha+\frac{1}{6} & \frac{1}{3}\alpha+\frac{1}{6} & \frac{1}{6}-\frac{1}{6}\alpha & \frac{1}{6}-\frac{1}{6}\alpha \\ \frac{1}{6}-\frac{1}{6}\alpha & \frac{1}{6} & \frac{1}{6}-\frac{1}{6}\alpha & \frac{1}{3}\alpha+\frac{1}{6} & \frac{1}{3}\alpha+\frac{1}{6} & \frac{1}{6}-\frac{1}{6}\alpha \end{bmatrix}$$

Add Teleportation
out of Dead Ends
(fills in empty cols)

Add Occasional Random
Teleportation to Also Escape Cycles

$$M = \alpha P' + (1-\alpha)\frac{1}{R}ee^T$$

Page Rank example – Final Google matrix

For $\alpha = 0.85$, we have:

$$M = \begin{bmatrix} \frac{1}{40} & \frac{1}{6} & \frac{37}{120} & \frac{1}{40} & \frac{1}{40} & \frac{1}{40} \\ \frac{9}{20} & \frac{1}{6} & \frac{37}{120} & \frac{1}{40} & \frac{1}{40} & \frac{1}{40} \\ \frac{9}{20} & \frac{1}{6} & \frac{1}{40} & \frac{1}{40} & \frac{1}{40} & \frac{1}{40} \\ \frac{1}{40} & \frac{1}{6} & \frac{1}{40} & \frac{1}{40} & \frac{9}{20} & \frac{7}{8} \\ \frac{1}{40} & \frac{1}{6} & \frac{37}{120} & \frac{9}{20} & \frac{1}{40} & \frac{1}{40} \\ \frac{1}{40} & \frac{1}{6} & \frac{1}{40} & \frac{9}{20} & \frac{9}{20} & \frac{1}{40} \end{bmatrix}.$$