$$\pi = (\alpha \mathrm{S} + (1 - \alpha)\mathrm{E})\pi$$

Sreejith, IIT Goa

The Setting

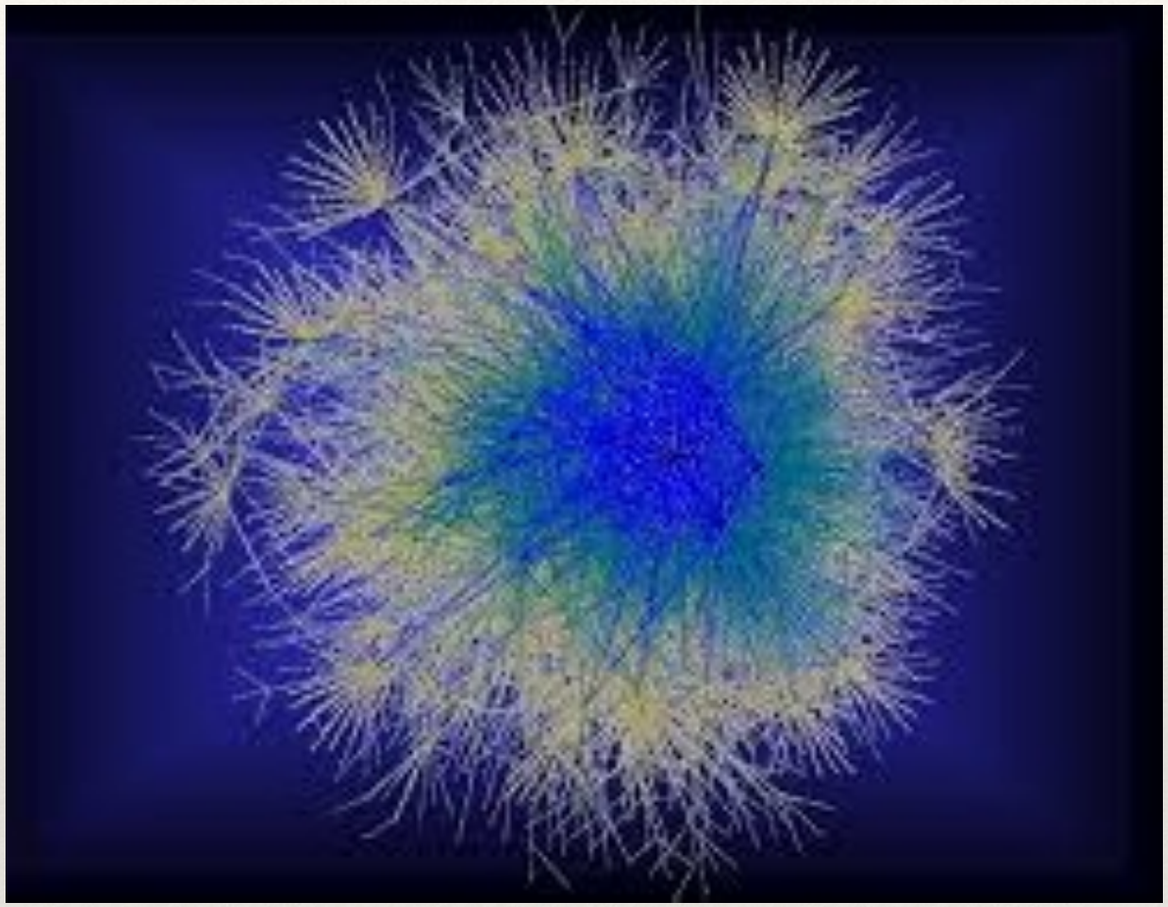
- I want to learn Chess.
- There are lakhs of websites which teach us to play Chess.
- But which one is the best? Which are the best 10 ones?
- This is the question, google pagerank tries to answer.
- Pagerank: Algorithm which ranks web-pages.
- Also means the ranking algorithm created by Larry Page, who along with Sergey Brin co-founded google.

Hyperlinks

- Consider two webpages which teach Chess:
  (A) Anand's page (B) Bran's page
- Let 5 pages have hyperlinks to A's page, and 1 page to B's
- More web pages interested in Anand's Chess than Bran's.
- Maybe, this is a good way to rank.
- What if Kasparov's page links to B's page whereas A's links are all from ordinary pages.
- The rank of the pages which link to A and B also important.
- **A page is important if it is pointed to by other important pages.**

Hyperlink Graph

- The vertices of the graph are web pages
- Edge from page X to page Y, if there is a hyperlink in X which points to Y.
- Number of webpages in the world = 1.5 billion



The Biggest Graph

First Attempt

- Observation: A page is important if it is pointed to by other pages.
- Let P_i be a page and $l(P_i)$ all pages which link to P_i
- Then, the rank of P_i is given by
- $$r(P_i) = \sum_{P_j \in l(P_i)} r(P_j)$$

First Attempt contd..

- Some webpages link to lots of pages, some to less.
- If there is one page which links to 100 Chess pages, whereas one page which links only to Anand's page, then values should differ.
- Similar to: A person's recommendation is more valuable if he/she gives less recommendations.

$$r(P_i) = \sum_{P_j \in l(P_i)} \frac{r(P_j)}{|P_j|}$$

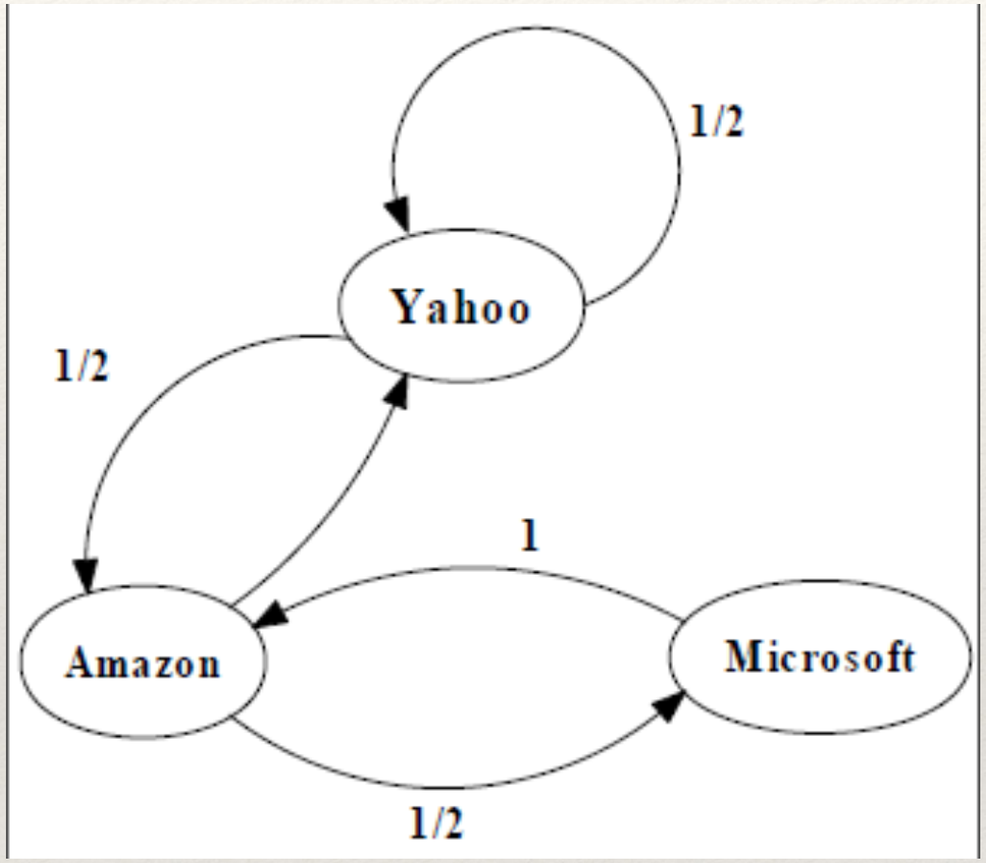
- $|P_j|$ denotes the number of hyperlinks in P_j

Computing rank

- We use this formula to compute the rank of all pages.
- At the beginning the ranks of all websites made equal.
- The equation is applied to computer the rank.
- The equation applied successively.

$$r_{k+1}(P_i) = \sum_{P_j \in l(P_i)} \frac{r_k(P_j)}{|P_j|}$$

Matrix Representation



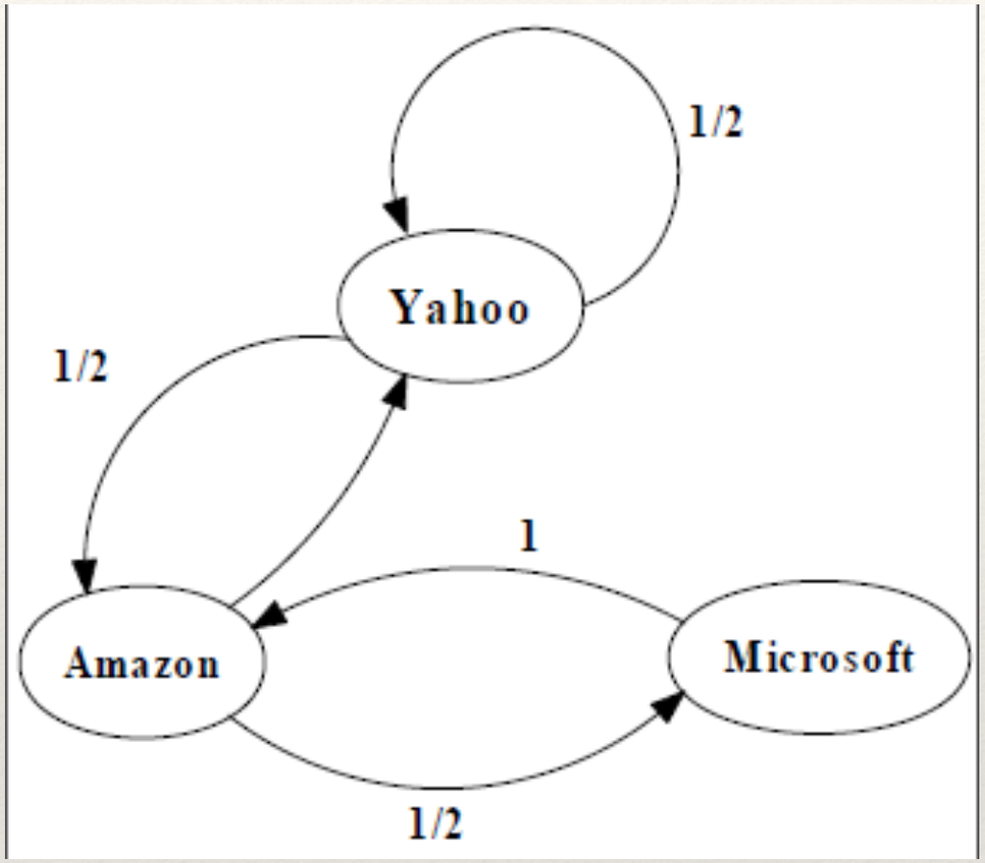
$$\begin{bmatrix} \text{yahoo} \\ \text{Amazon} \\ \text{Microsoft} \end{bmatrix} = \begin{bmatrix} 1/3 \\ 1/3 \\ 1/3 \end{bmatrix}$$

$$M = \begin{bmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 0 & 1 \\ 0 & 1/2 & 0 \end{bmatrix}$$

$$\begin{bmatrix} 1/3 \\ 1/2 \\ 1/6 \end{bmatrix} = \begin{bmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 0 & 1 \\ 0 & 1/2 & 0 \end{bmatrix} \begin{bmatrix} 1/3 \\ 1/3 \\ 1/3 \end{bmatrix}$$

copied from Fei Li's slides

Matrix Representation



$$\begin{bmatrix} 1/3 \\ 1/2 \\ 1/6 \end{bmatrix} = \begin{bmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 0 & 1 \\ 0 & 1/2 & 0 \end{bmatrix} \begin{bmatrix} 1/3 \\ 1/3 \\ 1/3 \end{bmatrix}$$

$$\begin{bmatrix} 5/12 \\ 1/3 \\ 1/4 \end{bmatrix} = \begin{bmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 0 & 1 \\ 0 & 1/2 & 0 \end{bmatrix} \begin{bmatrix} 1/3 \\ 1/2 \\ 1/6 \end{bmatrix}$$

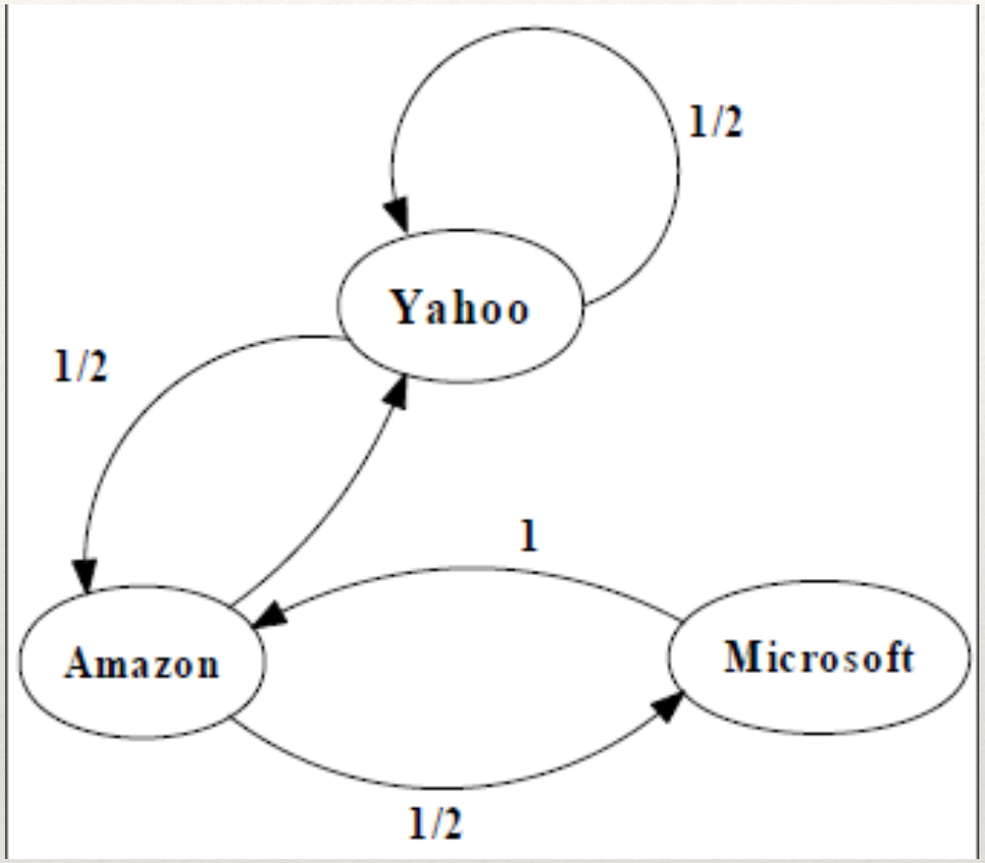
$$M = \begin{bmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 0 & 1 \\ 0 & 1/2 & 0 \end{bmatrix}$$

$$\begin{bmatrix} \text{yahoo} \\ \text{Amazon} \\ \text{Microsoft} \end{bmatrix} = \begin{bmatrix} 1/3 \\ 1/3 \\ 1/3 \end{bmatrix}$$

copied from Fei Li's slides

Pagerank = converging vector

copied from Fei Li's slides



$$\begin{bmatrix} 1/3 \\ 1/2 \\ 1/6 \end{bmatrix} = \begin{bmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 0 & 1 \\ 0 & 1/2 & 0 \end{bmatrix} \begin{bmatrix} 1/3 \\ 1/3 \\ 1/3 \end{bmatrix}$$

$$\begin{bmatrix} 5/12 \\ 1/3 \\ 1/4 \end{bmatrix} = \begin{bmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 0 & 1 \\ 0 & 1/2 & 0 \end{bmatrix} \begin{bmatrix} 1/3 \\ 1/2 \\ 1/6 \end{bmatrix}$$

$$M = \begin{bmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 0 & 1 \\ 0 & 1/2 & 0 \end{bmatrix}$$

$$\begin{bmatrix} \text{yahoo} \\ \text{Amazon} \\ \text{Microsoft} \end{bmatrix} = \begin{bmatrix} 1/3 \\ 1/3 \\ 1/3 \end{bmatrix}$$

$$\begin{bmatrix} 3/8 \\ 11/24 \\ 1/6 \end{bmatrix} \begin{bmatrix} 5/12 \\ 17/48 \\ 11/48 \end{bmatrix} \cdots \begin{bmatrix} 2/5 \\ 2/5 \\ 1/5 \end{bmatrix}$$

Converges

What's good in the Matrix?

- Each iteration involves a vector-matrix multiplication, which require $O(n^2)$ computation.
- The matrix is very sparse - most entries are 0. Estimates show, an average page has 10 links. Number of non-zero entries is = 10n.
- Sparse matrix multiplication can be done in O(n).
- Is M a Markov matrix?

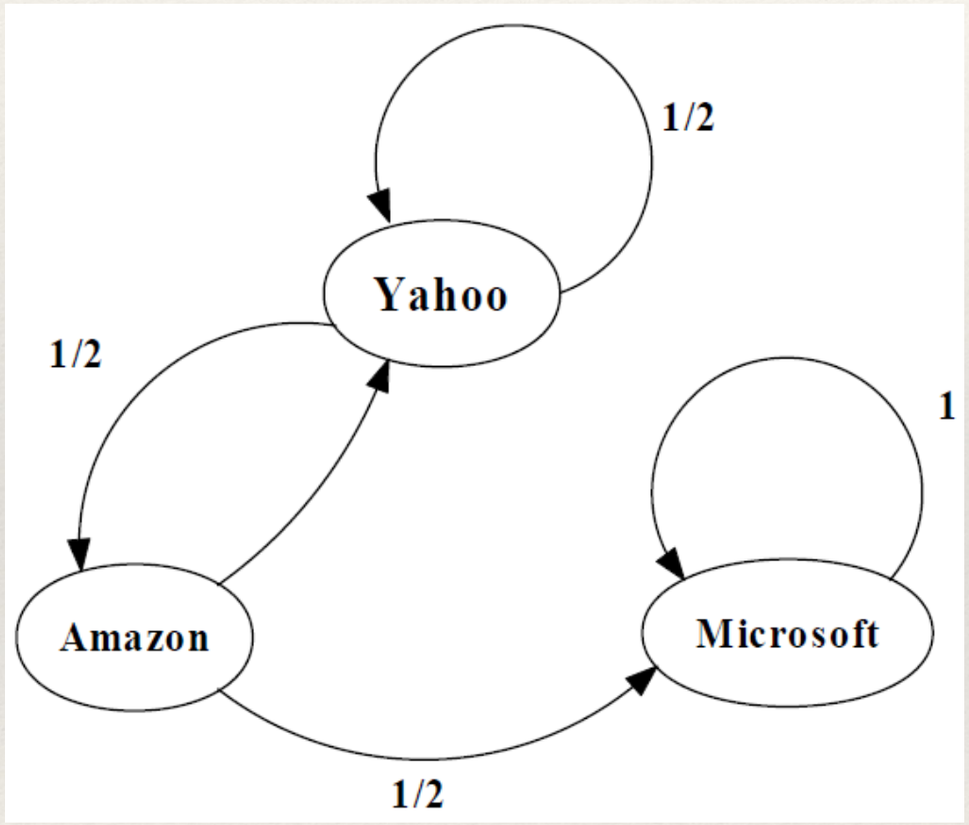
What's bad about the Matrix

- Will this rank computation go on indefinitely?
- Will this rank computation show periodic behaviour?
- Will it converge to multiple vectors?
- Does convergence depends on starting vector?
- Will convergence happen slowly?

Answer: Any of the above question can happen?

Another problem: loop

During each iteration, the loop accumulates rank but never distributes rank to other pages!



$$M = \begin{bmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 0 & 0 \\ 0 & 1/2 & 1 \end{bmatrix} \quad \begin{bmatrix} \text{yahoo} \\ \text{Amazon} \\ \text{Microsoft} \end{bmatrix} = \begin{bmatrix} 1/3 \\ 1/3 \\ 1/3 \end{bmatrix}$$

$$\begin{bmatrix} 1/3 \\ 1/6 \\ 1/2 \end{bmatrix} = \begin{bmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 0 & 0 \\ 0 & 1/2 & 1 \end{bmatrix} \begin{bmatrix} 1/3 \\ 1/3 \\ 1/3 \end{bmatrix}$$

$$\begin{bmatrix} 5/24 \\ 1/8 \\ 2/3 \end{bmatrix} \quad \begin{bmatrix} 1/6 \\ 5/48 \\ 35/48 \end{bmatrix} \quad .... \quad \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

Second attempt

- Random surfer model: A surfer goes to a page, clicks on a link randomly, and traverse the web (a random walk).
- If a page is repeated, the importance of the page increases.
- This is exactly the model we had with Matrix M.
- First problem: What to do, when we hit a page with no links?

Answer: Go to another random page. What changes required in matrix?

Markov Matrix

$$\begin{pmatrix} 0 & 0 & 1/3 & 0 & 0 & 0 \\ 1/2 & 0 & 1/3 & 0 & 0 & 0 \\ 1/2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1/2 & 1 \\ 0 & 0 & 1/3 & 1/2 & 0 & 0 \\ 0 & 0 & 0 & 1/2 & 1/2 & 0 \end{pmatrix} \dashrightarrow \begin{pmatrix} 0 & 1/6 & 1/3 & 0 & 0 & 0 \\ 1/2 & 1/6 & 1/3 & 0 & 0 & 0 \\ 1/2 & 1/6 & 0 & 0 & 0 & 0 \\ 0 & 1/6 & 0 & 0 & 1/2 & 1 \\ 0 & 1/6 & 1/3 & 1/2 & 0 & 0 \\ 0 & 1/6 & 0 & 1/2 & 1/2 & 0 \end{pmatrix}$$

M **S**

no links: 0 column vector

Markov chain: Sum of column is 1

$$S^T = M^T + \left(\frac{1}{n}\vec{1}\right)\vec{a}^T$$

Is this good enough?

Answer: No! no guarantee of convergence.

- Random surfer model: The surfer, walks through the web, but sometimes gets ``bored'' and randomly go to some other webpage and start walking from there.
- Gives the Google matrix

$$G = (\alpha S + (1 - \alpha)\frac{1}{n}\vec{1}\vec{1}^T)$$

$$\alpha = 0.85$$

Advantages of Google Matrix

$$G = (\alpha S + (1-\alpha)\frac{1}{n}\overrightarrow{1}\overrightarrow{1}^T)$$

- There is a unique converging vector for G. (because, all entries in G are strictly positive).

$$G\ \pi^* = \pi^*$$

- G is not sparse but still, computation can be done fast.

$$G\pi = (\alpha S + (1-\alpha)\frac{1}{n}\overrightarrow{1}\overrightarrow{1}^T)\pi$$

$$= \alpha M\pi + \left(\overrightarrow{1}(\alpha\overrightarrow{a}^T + (1-\alpha)\frac{1}{n}\overrightarrow{1}^T\right)\pi$$

- 20n steps, That is O(n) computation.

Pagerank = converging vector

- We check for convergence by repeatedly multiplying G.

$$\pi_1 = \mathrm{G}[1/n, 1/n, \ldots, 1/n]^T$$

$$\pi_2 = \mathrm{G}\pi_1$$

$$\vdots$$

$$\pi_{k+1} = \pi_k = \mathrm{G}\pi_k$$

- Total computation is = 20kn
- Page rank = π_k

Time taken: Depends on k

- G has an eigen value 1. Therefore, there exists eigen vector

$$\pi_k = G\pi_k$$

- The largest eigen value is 1 and second largest is α

$$G^{50}\pi \sim c_1.1^50.\pi^* + c_2\alpha^{50}\pi'$$

- For $\alpha = 0.85$, we have $\alpha^{50} = 0.000296$, good accuracy

- Total computation for pagerank: 20*50n steps = 1000n steps.

Thank you!