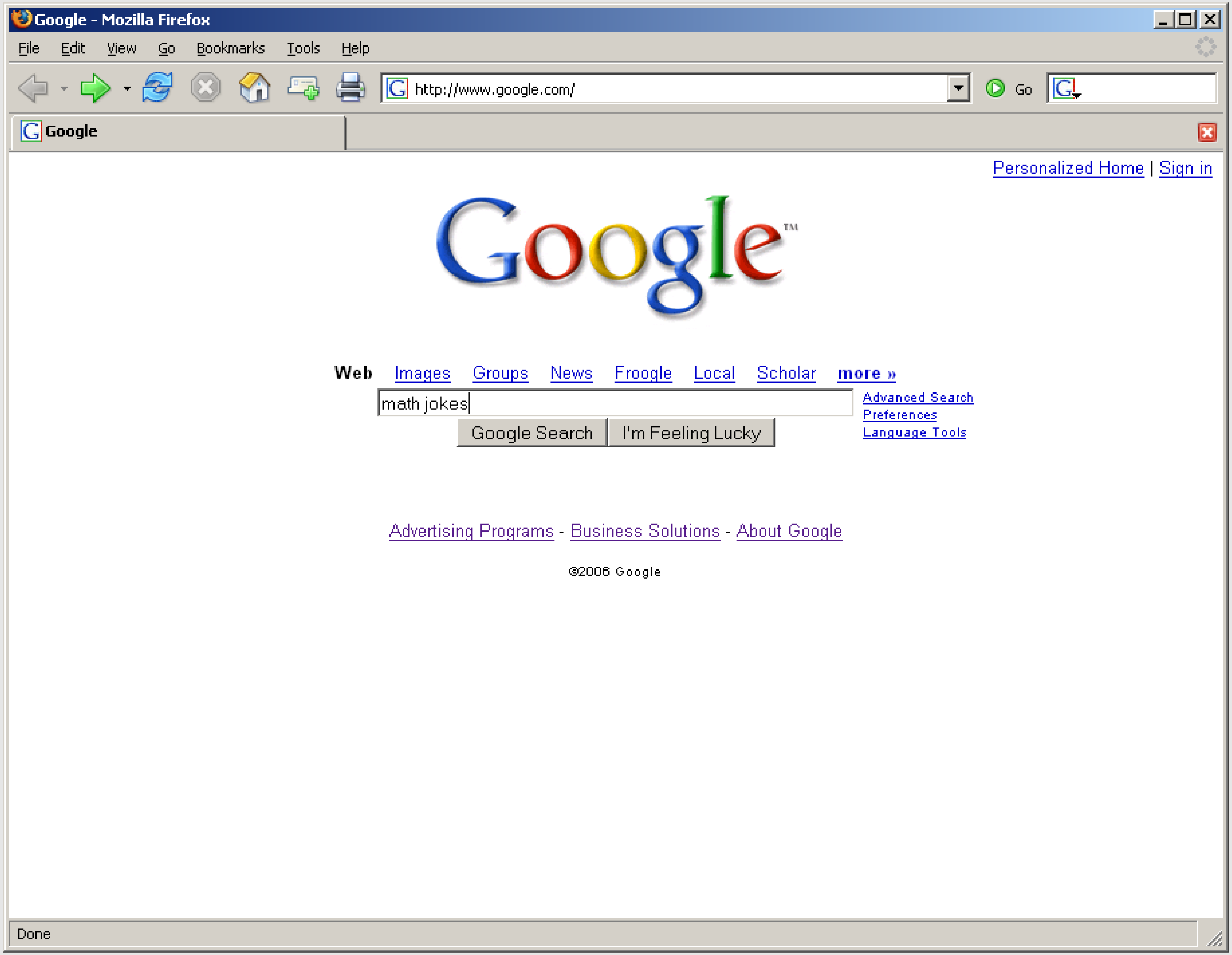
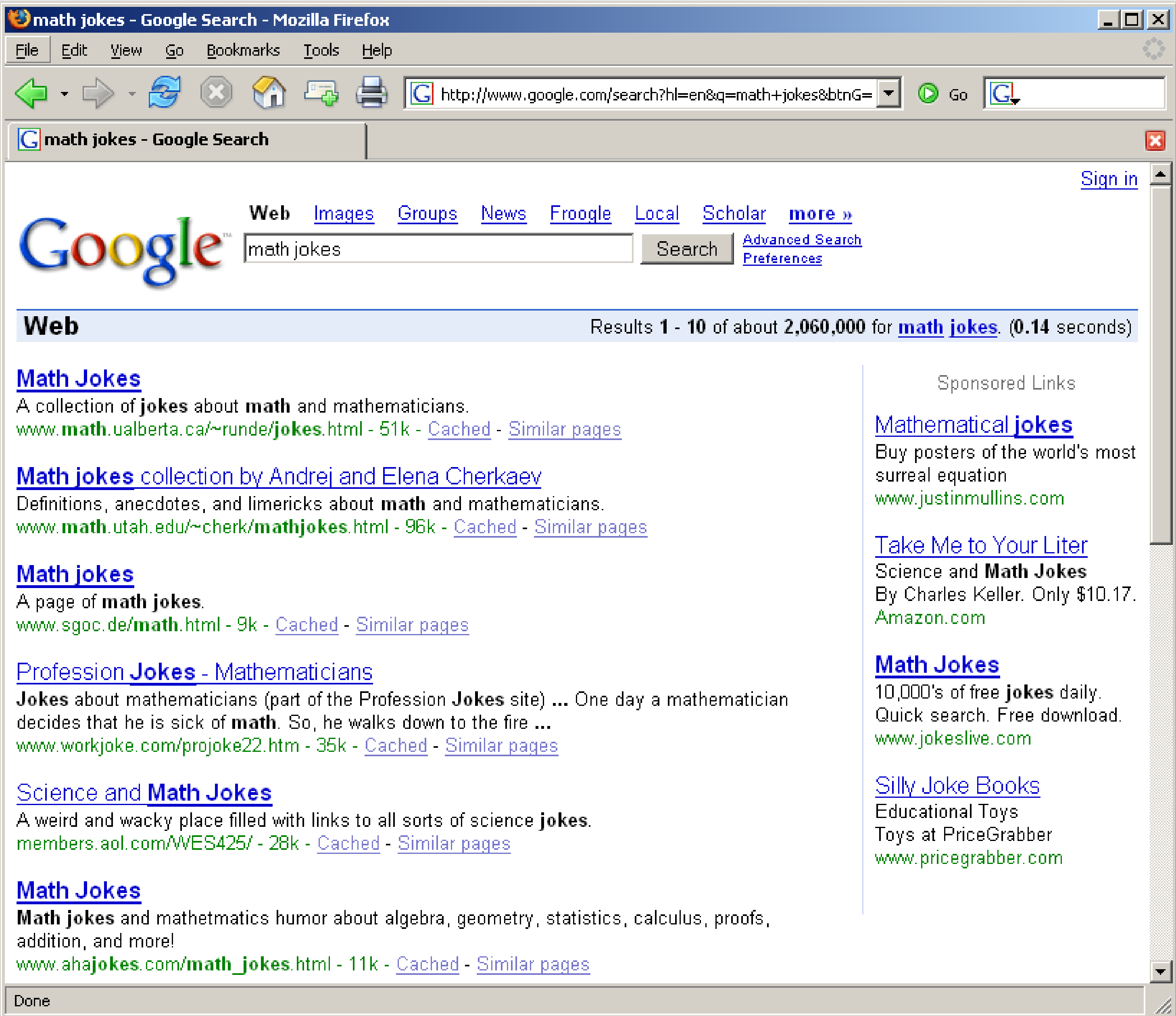
Probability, linear algebra, and numerical analysis: the mathematics behind Google's™ PageRank™

Grady Wright
Boise State University
wright@math.boisestate.edu

A Google search



A Google search



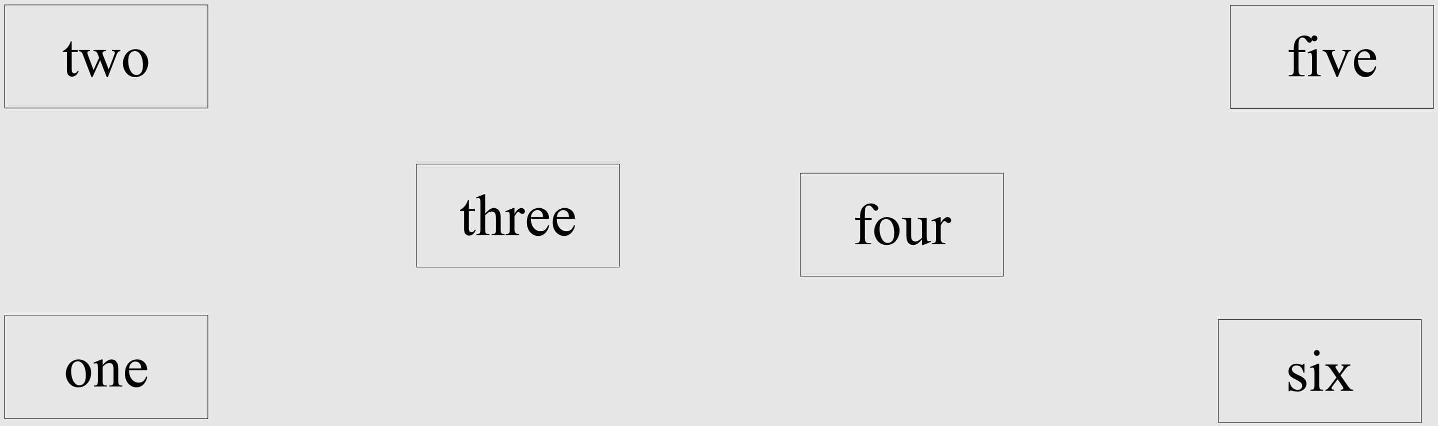
- Two step process
  1) Text processing
  2) Ranking
- Information Retrieval score
- PageRank™ score
  "Heart of Google software"
- Brin and Page (1998)
- Kleinberg: HITS
  www.teoma.com

Outline

- Heuristic interpretation of PageRank
- PageRank as a random walk (surf)
- Linear algebra formulation
- Computing PageRank
- Example and tools
- Advanced topics

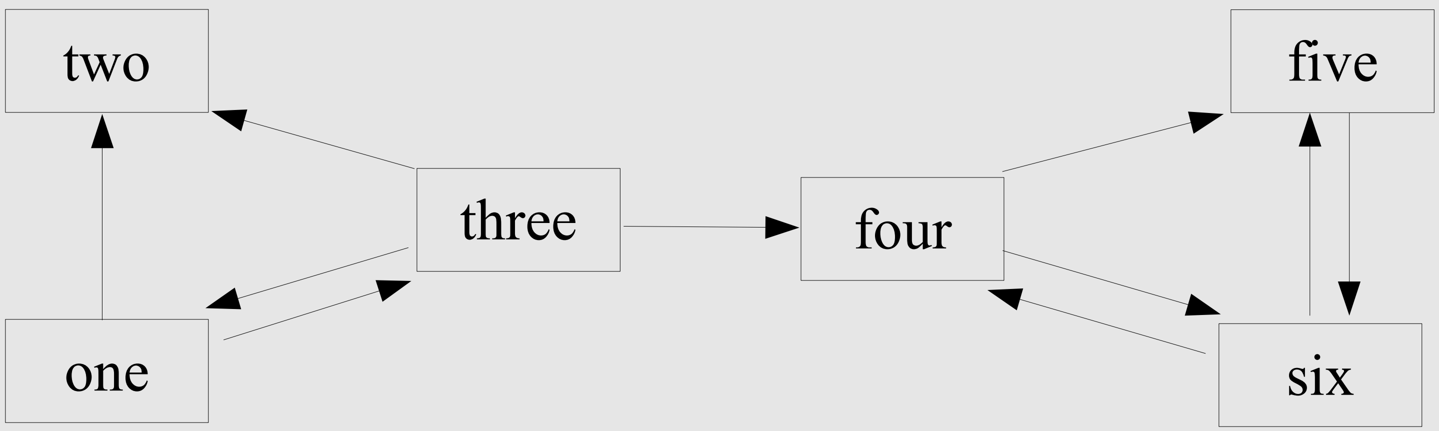
A tiny web example

- Pages of the web W:



A tiny web example

- Pages of the web *W* as a *directed graph*:

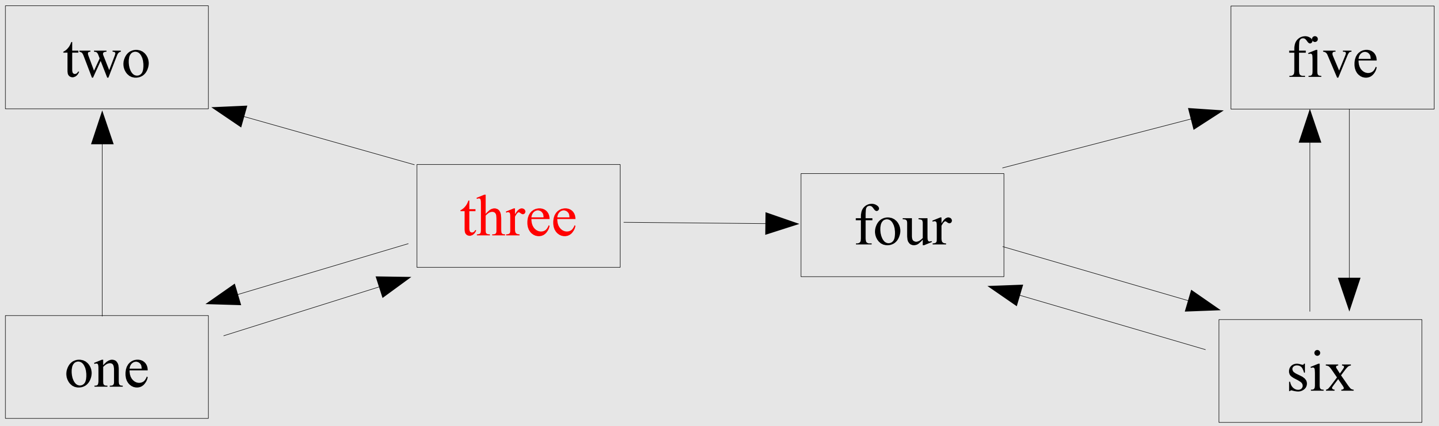


- Interpretation of PageRank:
    - A page is important if an important page has a link to it.
    - "Democracy of the web": a link from page A to page B is a vote from A to B.

- The *web according to* Google has over 11.5 billion pages (11,500,000,000)

PageRank: a random walk (or surf)

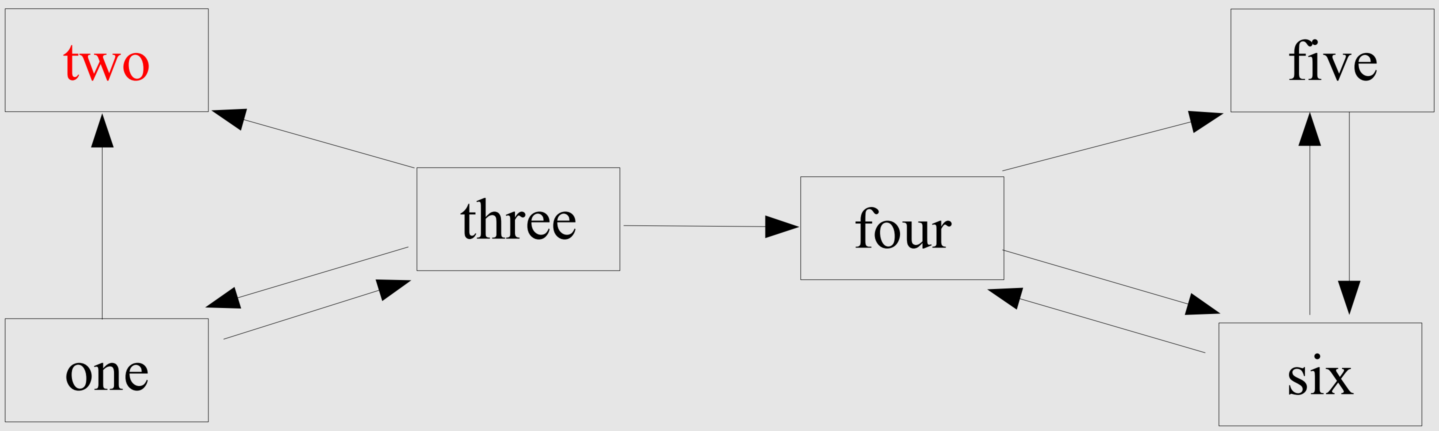
- Example:



- Infinitely dedicated random surfer
    - Outlinks

PageRank: a random walk (or surf)

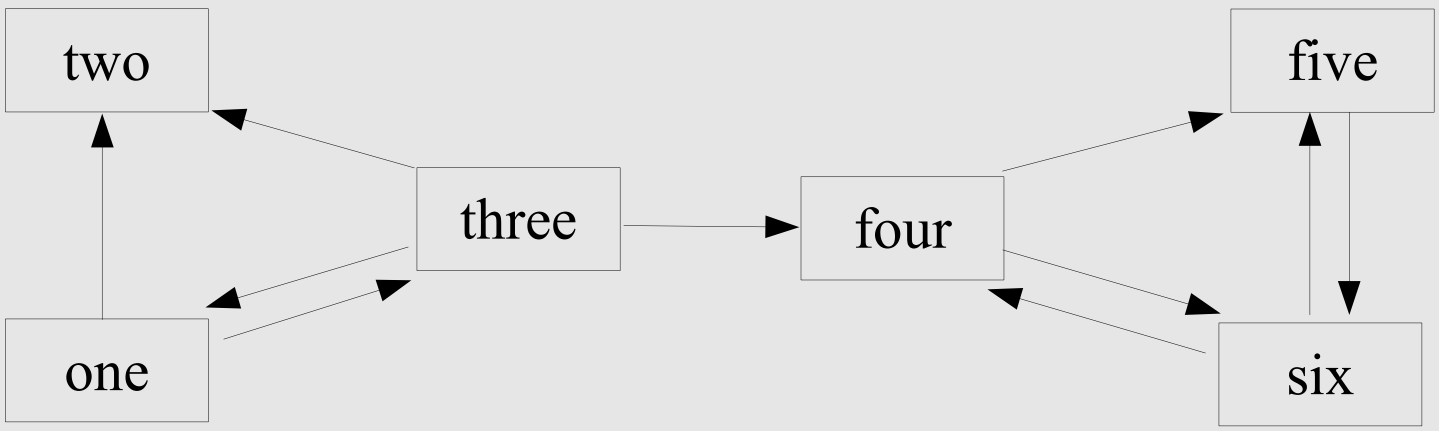
- Example:



- Infinitely dedicated random surfer
  - Outlinks
  - Dangling node

PageRank: a random walk (or surf)

- Example:



- Infinitely dedicated random surfer
  - Outlinks
  - Dangling node

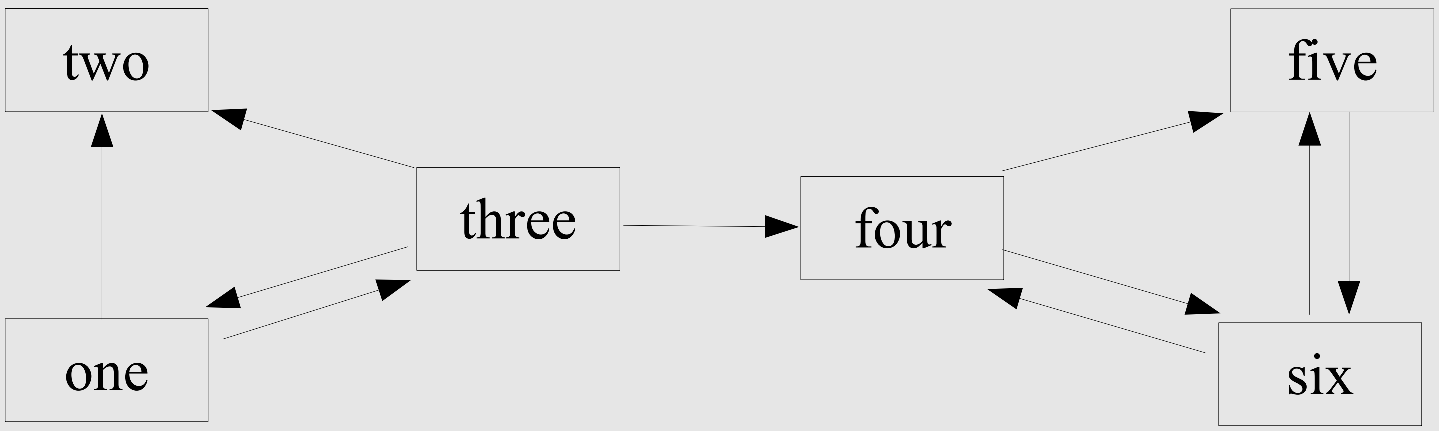
- Markov Chain

- Probabilistic interpretation of PageRank:

  A webpage's PageRank is the probability that at any particular time, the infinitely dedicated random surfer is visiting that page.

Linear algebra formulation

- Example:



- A directed graph can be represented using a *connectivity matrix* G.

<u>Entries of G:</u>

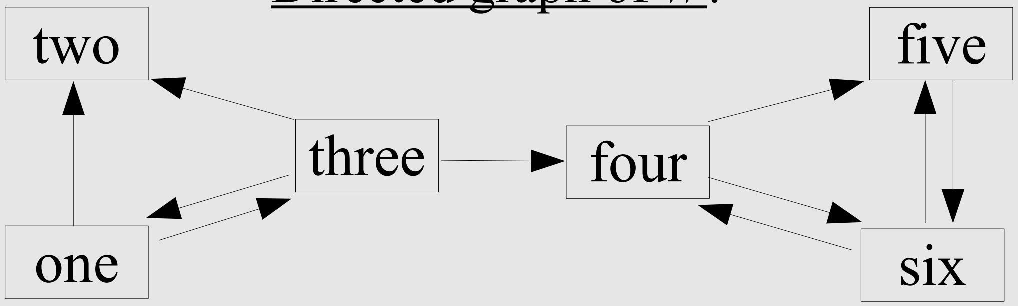
$$g_{i,j} = \begin{cases} 1 & \text{if page } j \text{ has a link to page } i \\ 0 & \text{otherwise} \end{cases} \quad \Rightarrow \quad G = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 \end{bmatrix}$$

$i,j = 1,2,\ldots,n$

Linear algebra formulation

- Add the random surf info to G:

Directed graph of W:



$$\vec{e} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}^T$$

Connectivity matrix:

$$G = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 \end{bmatrix}$$

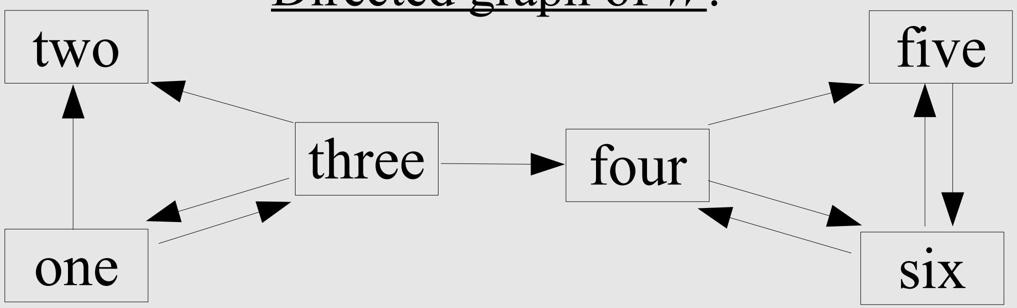
$$\vec{c} = \begin{bmatrix} 2 & 0 & 3 & 2 & 1 & 2 \end{bmatrix}^T$$

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

Linear algebra formulation

- Add the random surf info to G:

Directed graph of W:



$$\vec{e} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}^T$$

Connectivity matrix:

$$G = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 \end{bmatrix}$$

$$\vec{c} = \begin{bmatrix} 2 & 0 & 3 & 2 & 1 & 2 \end{bmatrix}^T$$

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

Transition probability matrix:

$$A = \underbrace{\begin{bmatrix} 0 & 0 & 1/3 & 0 & 0 & 0 \\ 1/2 & 0 & 1/3 & 0 & 0 & 0 \\ 1/2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1/3 & 0 & 0 & 1/2 \\ 0 & 0 & 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 1/2 & 1 & 0 \end{bmatrix}}_{P} + \underbrace{\begin{bmatrix} 0 & 1/6 & 0 & 0 & 0 & 0 \\ 0 & 1/6 & 0 & 0 & 0 & 0 \\ 0 & 1/6 & 0 & 0 & 0 & 0 \\ 0 & 1/6 & 0 & 0 & 0 & 0 \\ 0 & 1/6 & 0 & 0 & 0 & 0 \\ 0 & 1/6 & 0 & 0 & 0 & 0 \end{bmatrix}}_{\vec{e}\,\vec{d}^T/6} = \begin{bmatrix} 0 & 1/6 & 1/3 & 0 & 0 & 0 \\ 1/2 & 1/6 & 1/3 & 0 & 0 & 0 \\ 1/2 & 1/6 & 0 & 0 & 0 & 0 \\ 0 & 1/6 & 1/3 & 0 & 0 & 1/2 \\ 0 & 1/6 & 0 & 1/2 & 0 & 1/2 \\ 0 & 1/6 & 0 & 1/2 & 1 & 0 \end{bmatrix}$$

Linear algebra formulation

- For a general web W:

Define:

$$c_j = \sum_{i=1}^{n} g_{i,j} \quad j = 1,2,\ldots,n$$

(number of outgoing links from page j)

$$p_{i,j} = \begin{cases} \dfrac{g_{i,j}}{c_j} & \text{if } c_j \neq 0 \\ 0 & \text{otherwise} \end{cases} \quad i,j = 1,2,\ldots,n$$

(probability of visiting page i based on a random choice from the links on page j)

$$d_j = \begin{cases} 1 & \text{if } c_j = 0 \\ 0 & \text{otherwise} \end{cases} \quad j = 1,2,\ldots,n$$

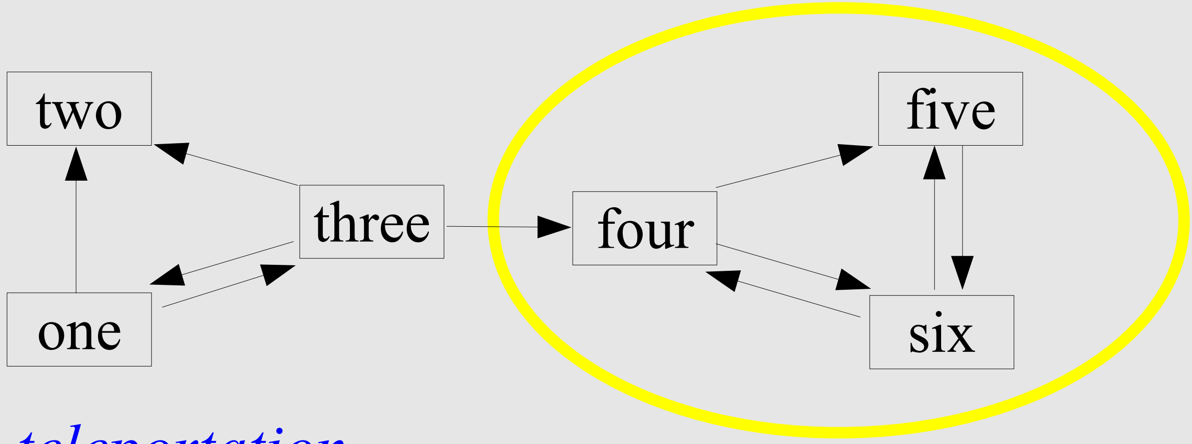
(tracks dangling pages)

- *Transition probability matrix A:*

$$A = P + \frac{\vec{e}}{n}\vec{d}^{\,T}, \text{ where } \vec{e} = \underbrace{\begin{bmatrix} 1 & 1 & \cdots & 1 \end{bmatrix}}_{n}^{T}$$

Avoiding cycles around cliques

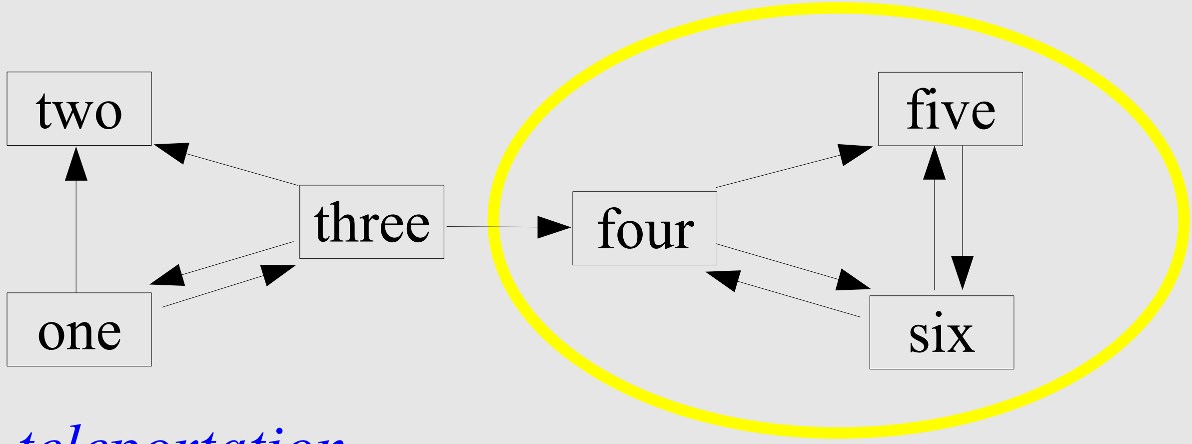
- Problem:



- Solution: *random teleportation*

Avoiding cycles around cliques

- Problem:



- Solution: *random teleportation*
- Modification to the *transition probability matrix*:

$$A = \alpha\left(P + \frac{\vec{e}}{n}\vec{d}^T\right) + (1-\alpha)\left(\frac{\vec{e}}{n}\vec{e}^T\right), \quad 0 < \alpha < 1$$

(Google originally set $\alpha = 0.85$)

- Matrix for above example:

$$A = \alpha\begin{bmatrix} 0 & 1/6 & 1/3 & 0 & 0 & 0 \\ 1/2 & 1/6 & 1/3 & 0 & 0 & 0 \\ 1/2 & 1/6 & 0 & 0 & 0 & 0 \\ 0 & 1/6 & 1/3 & 0 & 0 & 1/2 \\ 0 & 1/6 & 0 & 1/2 & 0 & 1/2 \\ 0 & 1/6 & 0 & 1/2 & 1 & 0 \end{bmatrix} + (1-\alpha)\begin{bmatrix} 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \end{bmatrix}$$

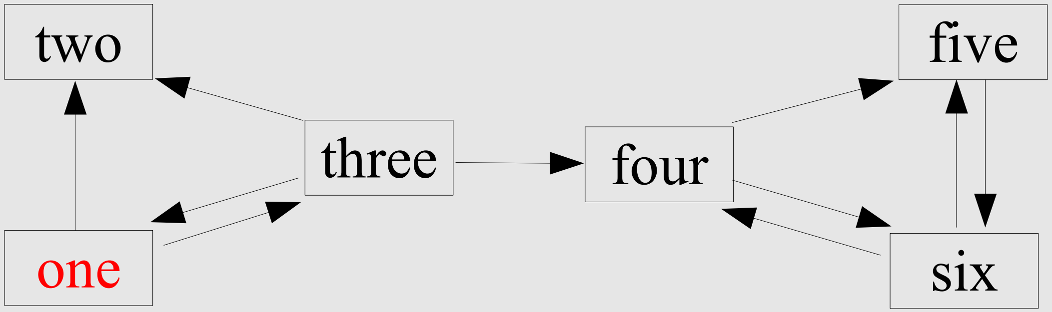
Importance of the transition prob. matrix

- *Probability distribution vector* $\vec{x}$ $\left(\sum_{j=1}^{n} x_j = 1\right)$:

  x_j = prob. the random surfer is currently visiting page *j*.

  $(A\vec{x})_j$ = prob. the random surfer will be visiting page *j* after leaving her current location.

- Example:
  $\alpha = 0.85$



$$\underbrace{\begin{bmatrix} 1/40 & 1/6 & 37/120 & 1/40 & 1/40 & 1/40 \\ 9/20 & 1/6 & 37/120 & 1/40 & 1/40 & 1/40 \\ 9/20 & 1/6 & 1/40 & 1/40 & 1/40 & 1/40 \\ 1/40 & 1/6 & 37/120 & 1/40 & 1/40 & 9/20 \\ 1/40 & 1/6 & 1/40 & 9/20 & 1/40 & 9/20 \\ 1/40 & 1/6 & 1/40 & 9/20 & 7/8 & 1/40 \end{bmatrix}}_{A} \underbrace{\begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}}_{x} = \begin{bmatrix} 0.025 \\ 0.45 \\ 0.45 \\ 0.025 \\ 0.025 \\ 0.025 \end{bmatrix}$$

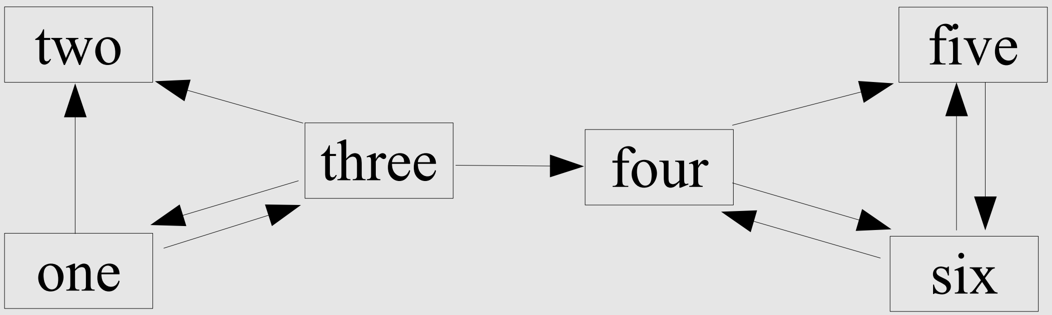
Importance of the transition prob. matrix

- *Probability distribution vector* $\vec{x}$ $\left(\sum_{j=1}^{n} x_j = 1\right)$:

  x_j = prob. the random surfer is currently visiting page j.

  $(A\vec{x})_j$ = prob. the random surfer will be visiting page j after leaving her current location.

- Example:
  $\alpha = 0.85$



$$\underbrace{\begin{bmatrix} 1/40 & 1/6 & 37/120 & 1/40 & 1/40 & 1/40 \\ 9/20 & 1/6 & 37/120 & 1/40 & 1/40 & 1/40 \\ 9/20 & 1/6 & 1/40 & 1/40 & 1/40 & 1/40 \\ 1/40 & 1/6 & 37/120 & 1/40 & 1/40 & 9/20 \\ 1/40 & 1/6 & 1/40 & 9/20 & 1/40 & 9/20 \\ 1/40 & 1/6 & 1/40 & 9/20 & 7/8 & 1/40 \end{bmatrix}}_{A} \underbrace{\begin{bmatrix} 0.025 \\ 0.45 \\ 0.45 \\ 0.025 \\ 0.025 \\ 0.025 \end{bmatrix}}_{x} = \begin{bmatrix} 0.21625 \\ 0.226875 \\ 0.099375 \\ 0.226875 \\ 0.11 \\ 0.120625 \end{bmatrix}$$

PageRank defined

- Page j's PageRank: j^{th} entry of the PDV $\vec{v}$ satisfying

$$\vec{v} = A\,\vec{v}$$

$\vec{v}$ = *stationary distribution vector* of A.

- What is the mathematical name for $\vec{v}$?

- Three concerns:

1. Existence

2. Uniqueness

3. Computation

Perron-Frobenius Theorem

Theorem: If A is an n-by-n matrix with positive entries then

1) One of its eigenvalues λ is positive and dominant.

2) There exists a unique (up to scaling) positive eigenvector corresponding to the dominant eigenvalue.

3) The dominant eigenvalue is simple.

Corollary: If the sum of each column of A equals 1 then $\lambda=1$ is the dominant eigenvalue.

Recall: $$A=\alpha P+\alpha\frac{\vec{e}}{n}\vec{d}^T+(1-\alpha)\frac{\vec{e}}{n}\vec{e}^T, \quad 0<\alpha<1$$

PageRank vector is the dominant eigenvector of A.

Computing PageRank: Power Method

- All we need is the dominant eigenvector!
- An idea:

$$\text{Define: } \overbrace{\{1, \lambda_2, \lambda_3, \ldots, \lambda_m\}, \{\vec{v}, \vec{v}_2, \vec{v}_3, \ldots, \vec{v}_m\}}^{\text{eigenvalues/vectors of } A} \quad (1 > |\lambda_2| \geq |\lambda_3| \geq \cdots \geq |\lambda_m|)$$

$$\text{Suppose: } \vec{x}^{(0)} = \vec{v} + \beta_2 \vec{v}_2 + \beta_3 \vec{v}_3 + \cdots + \beta_m \vec{v}_m \quad \left(\sum_{j=1}^{n} x_j^{(0)} = 1 \right)$$

$$\text{Consider: } \vec{x}^{(1)} = A \vec{x}^{(0)} = A \vec{v} + \beta_2 A \vec{v}_2 + \beta_3 A \vec{v}_3 + \cdots + \beta_m A \vec{v}_m$$

$$= \vec{v} + \beta_2 \lambda_2 \vec{v}_2 + \beta_3 \lambda_3 \vec{v}_3 + \cdots + \beta_m \lambda_m \vec{v}_m$$

Computing PageRank: Power Method

- All we need is the dominant eigenvector!

- An idea:

$$\text{Define: } \overbrace{\{1, \lambda_2, \lambda_3, \ldots, \lambda_m\}, \{\vec{v}, \vec{v}_2, \vec{v}_3, \ldots, \vec{v}_m\}}^{\text{eigenvalues/vectors of } A} \qquad (1 > |\lambda_2| \geq |\lambda_3| \geq \cdots \geq |\lambda_m|)$$

$$\text{Suppose: } \vec{x}^{(0)} = \vec{v} + \beta_2 \vec{v}_2 + \beta_3 \vec{v}_3 + \cdots + \beta_m \vec{v}_m \qquad \left(\sum_{j=1}^{n} x_j^{(0)} = 1\right)$$

$$\text{Consider: } \vec{x}^{(1)} = A\vec{x}^{(0)} = A\vec{v} + \beta_2 A\vec{v}_2 + \beta_3 A\vec{v}_3 + \cdots + \beta_m A\vec{v}_m$$

$$= \vec{v} + \beta_2 \lambda_2 \vec{v}_2 + \beta_3 \lambda_3 \vec{v}_3 + \cdots + \beta_m \lambda_m \vec{v}_m$$

$$\rightarrow \quad \vec{x}^{(2)} = A\vec{x}^{(1)} = A^2 \vec{x}^{(0)} = \vec{v} + \beta_2 \lambda_2^{\,2} \vec{v}_2 + \beta_3 \lambda_3^{\,2} \vec{v}_3 + \cdots + \beta_m \lambda_m^{\,2} \vec{v}_m$$

Computing PageRank: Power Method

- All we need is the dominant eigenvector!
- An idea:

Define: $\underbrace{\{1, \lambda_2, \lambda_3, \ldots, \lambda_m\}, \{\vec{v}, \vec{v}_2, \vec{v}_3, \ldots, \vec{v}_m\}}_{\text{eigenvalues/vectors of } A}$ $\quad (1 > |\lambda_2| \geq |\lambda_3| \geq \cdots \geq |\lambda_m|)$

Suppose: $\vec{x}^{(0)} = \vec{v} + \beta_2 \vec{v}_2 + \beta_3 \vec{v}_3 + \cdots + \beta_m \vec{v}_m \quad \left(\sum_{j=1}^{n} x_j^{(0)} = 1 \right)$

Consider: $\vec{x}^{(1)} = A \vec{x}^{(0)} = A \vec{v} + \beta_2 A \vec{v}_2 + \beta_3 A \vec{v}_3 + \cdots + \beta_m A \vec{v}_m$

$$= \vec{v} + \beta_2 \lambda_2 \vec{v}_2 + \beta_3 \lambda_3 \vec{v}_3 + \cdots + \beta_m \lambda_m \vec{v}_m$$

$$\vec{x}^{(2)} = A \vec{x}^{(1)} = A^2 \vec{x}^{(0)} = \vec{v} + \beta_2 \lambda_2^{\,2} \vec{v}_2 + \beta_3 \lambda_3^{\,2} \vec{v}_3 + \cdots + \beta_m \lambda_m^{\,2} \vec{v}_m$$

$$\vdots$$

$$\rightarrow \quad \vec{x}^{(k+1)} = A \vec{x}^{(k)} = A^k \vec{x}^{(0)} = \vec{v} + \beta_2 \lambda_2^{\,k} \vec{v}_2 + \beta_3 \lambda_3^{\,k} \vec{v}_3 + \cdots + \beta_m \lambda_m^{\,k} \vec{v}_m$$

Computing PageRank: Power Method

- All we need is the dominant eigenvector!
- An idea:

$$\text{Define: } \overbrace{\{1, \lambda_2, \lambda_3, \ldots, \lambda_m\}, \{\vec{v}, \vec{v}_2, \vec{v}_3, \ldots, \vec{v}_m\}}^{\text{eigenvalues/vectors of } A} \qquad (1 > |\lambda_2| \geq |\lambda_3| \geq \cdots \geq |\lambda_m|)$$

$$\text{Suppose: } \vec{x}^{(0)} = \vec{v} + \beta_2 \vec{v}_2 + \beta_3 \vec{v}_3 + \cdots + \beta_m \vec{v}_m \qquad \left(\sum_{j=1}^{n} x_j^{(0)} = 1 \right)$$

$$\text{Consider: } \vec{x}^{(1)} = A\vec{x}^{(0)} = A\vec{v} + \beta_2 A\vec{v}_2 + \beta_3 A\vec{v}_3 + \cdots + \beta_m A\vec{v}_m$$

$$= \vec{v} + \beta_2 \lambda_2 \vec{v}_2 + \beta_3 \lambda_3 \vec{v}_3 + \cdots + \beta_m \lambda_m \vec{v}_m$$

$$\vec{x}^{(2)} = A\vec{x}^{(1)} = A^2 \vec{x}^{(0)} = \vec{v} + \beta_2 \lambda_2^{\,2} \vec{v}_2 + \beta_3 \lambda_3^{\,2} \vec{v}_3 + \cdots + \beta_m \lambda_m^{\,2} \vec{v}_m$$

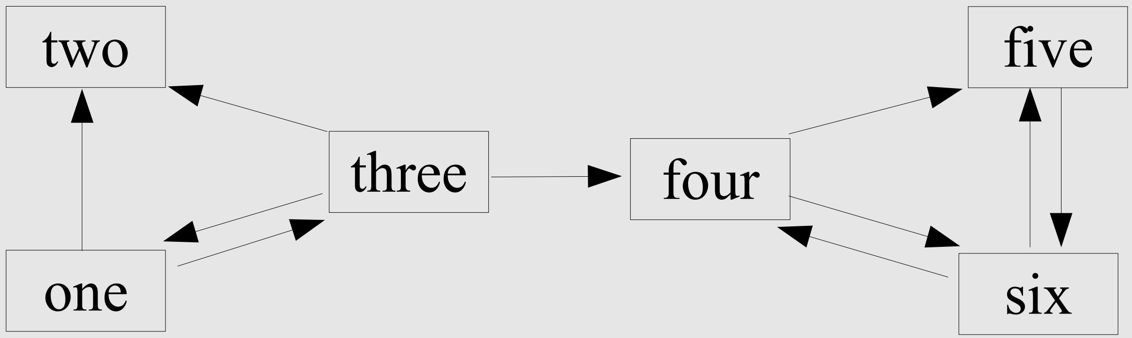
$$\vdots$$

$$\vec{x}^{(k+1)} = A\vec{x}^{(k)} = A^k \vec{x}^{(0)} = \vec{v} + \beta_2 \lambda_2^{\,k} \vec{v}_2 + \beta_3 \lambda_3^{\,k} \vec{v}_3 + \cdots + \beta_m \lambda_m^{\,k} \vec{v}_m$$

Thus: $\vec{x}^{(k)}$ converges to the PageRank vector $\vec{v}$ as $k \to \infty$.

Example of power method

- Example:
  $\alpha = 0.85$



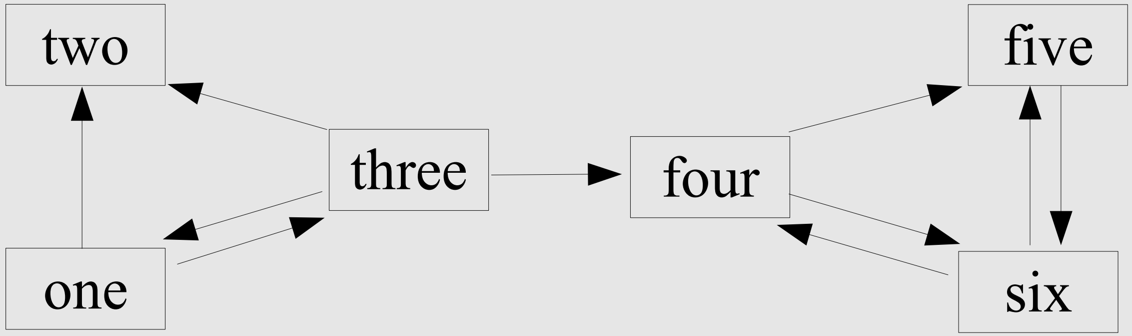
- Initial guess: $\vec{x}^{(0)} = \begin{bmatrix} 1 & 1 & \cdots & 1 \end{bmatrix}^T / n$

$$\underbrace{\begin{bmatrix} 0.095833333 \\ 0.16666667 \\ 0.11944444 \\ 0.16666667 \\ 0.19027778 \\ 0.26111111 \end{bmatrix}}_{\vec{x}^{(1)}} = \begin{bmatrix} 1/40 & 1/6 & 37/120 & 1/40 & 1/40 & 1/40 \\ 9/20 & 1/6 & 37/120 & 1/40 & 1/40 & 1/40 \\ 9/20 & 1/6 & 1/40 & 1/40 & 1/40 & 1/40 \\ 1/40 & 1/6 & 37/120 & 1/40 & 1/40 & 9/20 \\ 1/40 & 1/6 & 1/40 & 9/20 & 1/40 & 9/20 \\ 1/40 & 1/6 & 1/40 & 9/20 & 7/8 & 1/40 \end{bmatrix} \underbrace{\begin{bmatrix} 1/6 \\ 1/6 \\ 1/6 \\ 1/6 \\ 1/6 \\ 1/6 \end{bmatrix}}_{\vec{x}^{(0)}}$$

$$\|\vec{x}^{(1)} - \vec{x}^{(0)}\|_\infty = max_{1 \le j \le n} |x_j^{(1)} - x_j^{(0)}| = 9.4 \cdot 10^{-1}$$

Example of power method

- Example:
  $\alpha = 0.85$



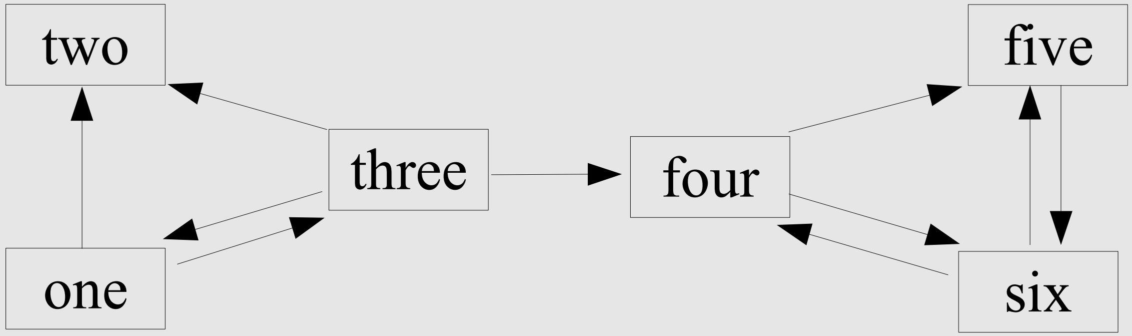
- Initial guess: $\vec{x}^{(0)} = \begin{bmatrix} 1 & 1 & \cdots & 1 \end{bmatrix}^T / n$

$$\underbrace{\begin{bmatrix} 0.082453704 \\ 0.12318287 \\ 0.089340278 \\ 0.19342593 \\ 0.23041667 \\ 0.28118056 \end{bmatrix}}_{\vec{x}^{(2)}} = \begin{bmatrix} 1/40 & 1/6 & 37/120 & 1/40 & 1/40 & 1/40 \\ 9/20 & 1/6 & 37/120 & 1/40 & 1/40 & 1/40 \\ 9/20 & 1/6 & 1/40 & 1/40 & 1/40 & 1/40 \\ 1/40 & 1/6 & 37/120 & 1/40 & 1/40 & 9/20 \\ 1/40 & 1/6 & 1/40 & 9/20 & 1/40 & 9/20 \\ 1/40 & 1/6 & 1/40 & 9/20 & 7/8 & 1/40 \end{bmatrix} \underbrace{\begin{bmatrix} 0.095833333 \\ 0.16666667 \\ 0.11944444 \\ 0.16666667 \\ 0.19027778 \\ 0.26111111 \end{bmatrix}}_{\vec{x}^{(1)}}$$

$$\|\vec{x}^{(2)} - \vec{x}^{(1)}\|_\infty = 4.3 \cdot 10^{-2}$$

Example of power method

- Example:
  $\alpha = 0.85$



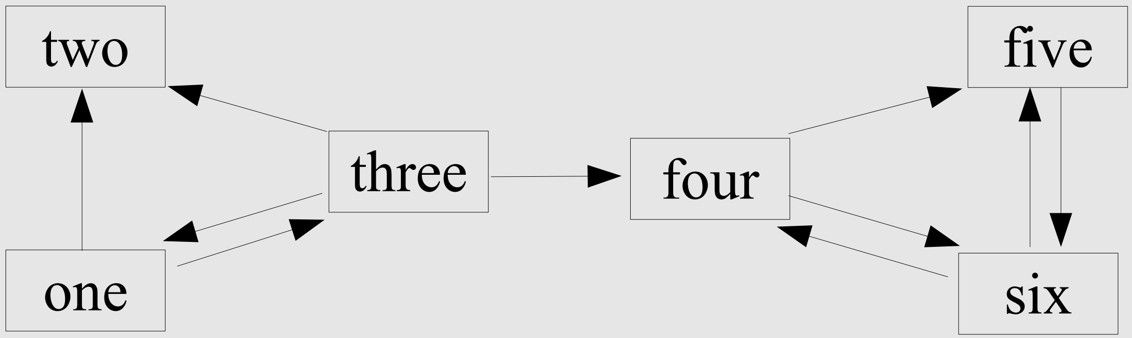
- Initial guess: $\vec{x}^{(0)} = \begin{bmatrix} 1 & 1 & \cdots & 1 \end{bmatrix}^T / n$

$$\underbrace{\begin{bmatrix} 0.067763985 \\ 0.10280681 \\ 0.077493731 \\ 0.18722657 \\ 0.24415866 \\ 0.32051109 \end{bmatrix}}_{\vec{x}^{(3)}} = \begin{bmatrix} 1/40 & 1/6 & 37/120 & 1/40 & 1/40 & 1/40 \\ 9/20 & 1/6 & 37/120 & 1/40 & 1/40 & 1/40 \\ 9/20 & 1/6 & 1/40 & 1/40 & 1/40 & 1/40 \\ 1/40 & 1/6 & 37/120 & 1/40 & 1/40 & 9/20 \\ 1/40 & 1/6 & 1/40 & 9/20 & 1/40 & 9/20 \\ 1/40 & 1/6 & 1/40 & 9/20 & 7/8 & 1/40 \end{bmatrix} \underbrace{\begin{bmatrix} 0.082453704 \\ 0.12318287 \\ 0.089340278 \\ 0.19342593 \\ 0.23041667 \\ 0.28118056 \end{bmatrix}}_{\vec{x}^{(2)}}$$

$$\|\vec{x}^{(3)} - \vec{x}^{(2)}\|_\infty = 3.9 \cdot 10^{-2}$$

Example of power method

- Example:
  $\alpha = 0.85$



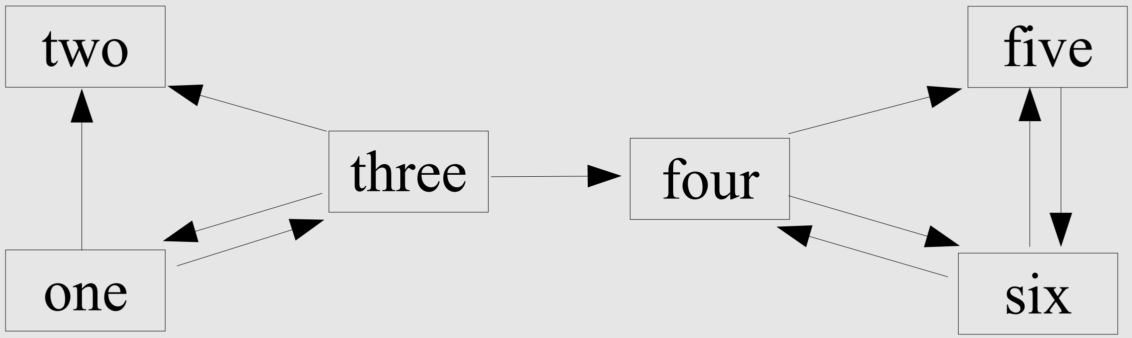
- Initial guess: $\vec{x}^{(0)} = \begin{bmatrix} 1 & 1 & \cdots & 1 \end{bmatrix}^T / n$

$$\underbrace{\begin{bmatrix} 0.061520855 \\ 0.090320549 \\ 0.068363992 \\ 0.19773807 \\ 0.25536944 \\ 0.32668709 \end{bmatrix}}_{\vec{x}^{(4)}} = \begin{bmatrix} 1/40 & 1/6 & 37/120 & 1/40 & 1/40 & 1/40 \\ 9/20 & 1/6 & 37/120 & 1/40 & 1/40 & 1/40 \\ 9/20 & 1/6 & 1/40 & 1/40 & 1/40 & 1/40 \\ 1/40 & 1/6 & 37/120 & 1/40 & 1/40 & 9/20 \\ 1/40 & 1/6 & 1/40 & 9/20 & 1/40 & 9/20 \\ 1/40 & 1/6 & 1/40 & 9/20 & 7/8 & 1/40 \end{bmatrix} \underbrace{\begin{bmatrix} 0.067763985 \\ 0.10280681 \\ 0.077493731 \\ 0.18722657 \\ 0.24415866 \\ 0.32051109 \end{bmatrix}}_{\vec{x}^{(3)}}$$

$$\|\vec{x}^{(4)} - \vec{x}^{(3)}\|_\infty = 1.3 \cdot 10^{-2}$$

Example of power method

- Example:
  $\alpha = 0.85$



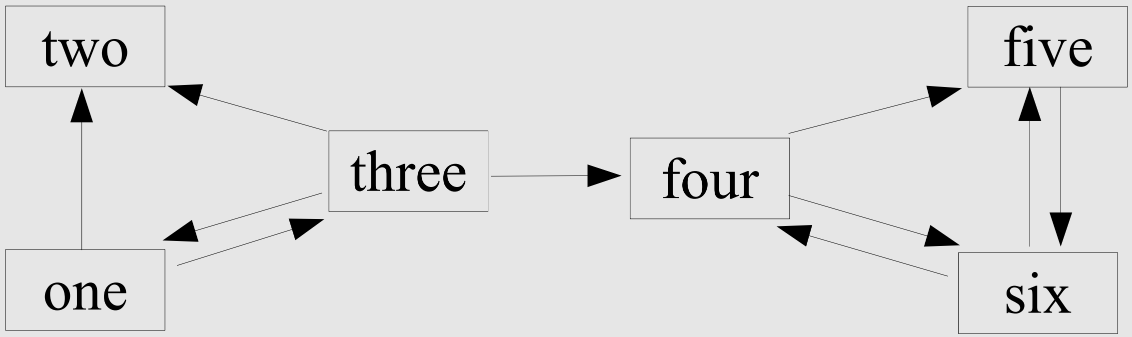
- Initial guess: $\vec{x}^{(0)} = \begin{bmatrix} 1 & 1 & \cdots & 1 \end{bmatrix}^T / n$

$$\underbrace{\begin{bmatrix} 0.057165209 \\ 0.083311572 \\ 0.063941774 \\ 0.19600722 \\ 0.26067610 \\ 0.33889812 \end{bmatrix}}_{\vec{x}^{(5)}} = \begin{bmatrix} 1/40 & 1/6 & 37/120 & 1/40 & 1/40 & 1/40 \\ 9/20 & 1/6 & 37/120 & 1/40 & 1/40 & 1/40 \\ 9/20 & 1/6 & 1/40 & 1/40 & 1/40 & 1/40 \\ 1/40 & 1/6 & 37/120 & 1/40 & 1/40 & 9/20 \\ 1/40 & 1/6 & 1/40 & 9/20 & 1/40 & 9/20 \\ 1/40 & 1/6 & 1/40 & 9/20 & 7/8 & 1/40 \end{bmatrix} \underbrace{\begin{bmatrix} 0.061520855 \\ 0.090320549 \\ 0.068363992 \\ 0.19773807 \\ 0.25536944 \\ 0.32668709 \end{bmatrix}}_{\vec{x}^{(4)}}$$

$$\|\vec{x}^{(5)} - \vec{x}^{(4)}\|_\infty = 1.2 \cdot 10^{-2}$$

Example of power method

- Example:
  $\alpha = 0.85$



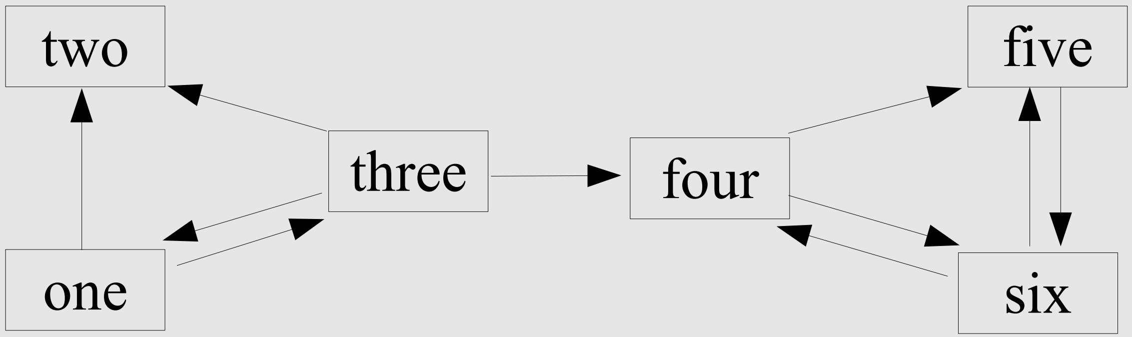
- Initial guess: $\vec{x}^{(0)} = \begin{bmatrix} 1 & 1 & \cdots & 1 \end{bmatrix}^T / n$

$$\underbrace{\begin{bmatrix} 0.054919309 \\ 0.079214523 \\ 0.061097686 \\ 0.19895101 \\ 0.26413724 \\ 0.34168023 \end{bmatrix}}_{\vec{x}^{(6)}} = \begin{bmatrix} 1/40 & 1/6 & 37/120 & 1/40 & 1/40 & 1/40 \\ 9/20 & 1/6 & 37/120 & 1/40 & 1/40 & 1/40 \\ 9/20 & 1/6 & 1/40 & 1/40 & 1/40 & 1/40 \\ 1/40 & 1/6 & 37/120 & 1/40 & 1/40 & 9/20 \\ 1/40 & 1/6 & 1/40 & 9/20 & 1/40 & 9/20 \\ 1/40 & 1/6 & 1/40 & 9/20 & 7/8 & 1/40 \end{bmatrix} \underbrace{\begin{bmatrix} 0.057165209 \\ 0.083311572 \\ 0.063941774 \\ 0.19600722 \\ 0.26067610 \\ 0.33889812 \end{bmatrix}}_{\vec{x}^{(5)}}$$

$$\|\vec{x}^{(6)} - \vec{x}^{(5)}\|_\infty = 4.1 \cdot 10^{-3}$$

Example of power method

- Example:
  $\alpha = 0.85$



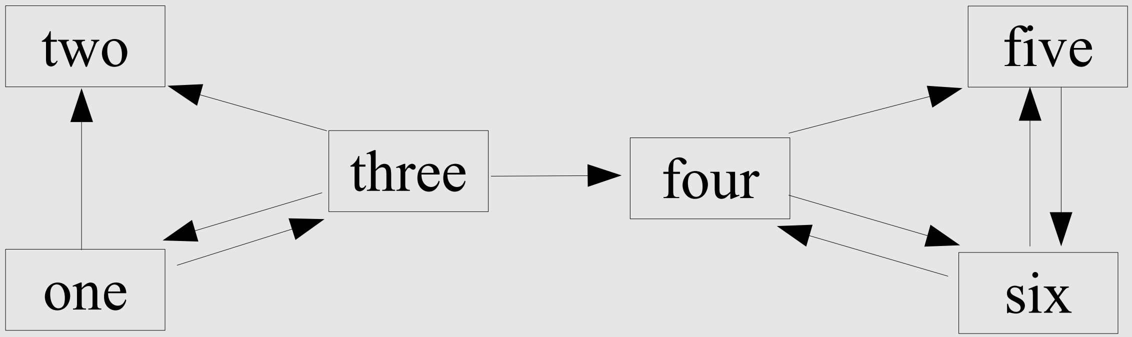
- Initial guess: $\vec{x}^{(0)} = \begin{bmatrix} 1 & 1 & \cdots & 1 \end{bmatrix}^T / n$

$$\underbrace{\begin{bmatrix} 0.053533069 \\ 0.076873775 \\ 0.059562764 \\ 0.19874717 \\ 0.26599033 \\ 0.34529289 \end{bmatrix}}_{\vec{x}^{(7)}} = \begin{bmatrix} 1/40 & 1/6 & 37/120 & 1/40 & 1/40 & 1/40 \\ 9/20 & 1/6 & 37/120 & 1/40 & 1/40 & 1/40 \\ 9/20 & 1/6 & 1/40 & 1/40 & 1/40 & 1/40 \\ 1/40 & 1/6 & 37/120 & 1/40 & 1/40 & 9/20 \\ 1/40 & 1/6 & 1/40 & 9/20 & 1/40 & 9/20 \\ 1/40 & 1/6 & 1/40 & 9/20 & 7/8 & 1/40 \end{bmatrix} \underbrace{\begin{bmatrix} 0.054919309 \\ 0.079214523 \\ 0.061097686 \\ 0.19895101 \\ 0.26413724 \\ 0.34168023 \end{bmatrix}}_{\vec{x}^{(6)}}$$

$$\|\vec{x}^{(6)} - \vec{x}^{(5)}\|_\infty = 3.6 \cdot 10^{-3}$$

Example of power method

- Example:
  $\alpha = 0.85$



- Initial guess: $\vec{x}^{(0)} = \begin{bmatrix} 1 & 1 & \cdots & 1 \end{bmatrix}^T / n$

$$\underbrace{\begin{bmatrix} 0.051704746 \\ 0.073679263 \\ 0.057412413 \\ 0.19990381 \\ 0.26859608 \\ 0.34870368 \end{bmatrix}}_{\vec{x}^{(35)}} = \begin{bmatrix} 1/40 & 1/6 & 37/120 & 1/40 & 1/40 & 1/40 \\ 9/20 & 1/6 & 37/120 & 1/40 & 1/40 & 1/40 \\ 9/20 & 1/6 & 1/40 & 1/40 & 1/40 & 1/40 \\ 1/40 & 1/6 & 37/120 & 1/40 & 1/40 & 9/20 \\ 1/40 & 1/6 & 1/40 & 9/20 & 1/40 & 9/20 \\ 1/40 & 1/6 & 1/40 & 9/20 & 7/8 & 1/40 \end{bmatrix} \underbrace{\begin{bmatrix} 0.051704746 \\ 0.073679264 \\ 0.057412413 \\ 0.19990381 \\ 0.26859608 \\ 0.34870368 \end{bmatrix}}_{\vec{x}^{(34)}}$$

$\approx$ PageRank

$$\|\vec{x}^{(35)} - \vec{x}^{(34)}\|_\infty = 1.0 \cdot 10^{-9}$$

Power method algorithm and efficiency

- Power method algorithm:

$$\vec{x}^{(0)} = \begin{bmatrix} 1 & 1 & \cdots & 1 \end{bmatrix}^T / n$$
$$k = 1$$
repeat
$$\vec{x}^{(k)} = A\vec{x}^{(k-1)}$$
$$\delta = \| \vec{x}^{(k)} - \vec{x}^{(k-1)} \|_\infty$$
$$k = k + 1$$
until $\delta \le \epsilon$

Cocktail napkin computational cost analysis for Google

Operation	# FLOPs
$A\vec{x}^{(k-1)}$	$2n^2 - n$
$\vec{x}^{(k)} - \vec{x}^{(k-1)}$	n
Total	$2n^2 = O(n^2)$

- $n = 11.5 \cdot 10^9$
- FLOPs $\approx 2.65 \cdot 10^{20}$
- IBM BlueGene/L: $478 \cdot 10^{12}$ FLOPs/sec
- 1 iteration: 154 hours
- 50 – 100 iterations: 45 – 91 weeks!

More efficient power method for PageRank

- Idea: exploit the structure of the *transition probability matrix*

Recall: $$A = \alpha P + \alpha \frac{\vec{e}}{n} \vec{d}^T + (1-\alpha) \frac{\vec{e}}{n} \vec{e}^T, \quad 0 < \alpha < 1$$

Thus: $$\vec{x}^{(k)} = A \vec{x}^{(k-1)} = \alpha P \vec{x}^{(k-1)} + \alpha \frac{\vec{e}}{n} \vec{d}^T \vec{x}^{(k-1)} + (1-\alpha) \frac{\vec{e}}{n} \vec{e}^T \vec{x}^{(k-1)}$$

$$\vec{x}^{(k)} = \alpha P \vec{x}^{(k-1)} + \frac{\vec{e}}{n} \left(\alpha \vec{d}^T \vec{x}^{(k-1)} + (1-\alpha) \right)$$

- Example:

$$\vec{x}^{(k)} = \alpha \begin{bmatrix} 0 & 0 & 1/3 & 0 & 0 & 0 \\ 1/2 & 0 & 1/3 & 0 & 0 & 0 \\ 1/2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1/3 & 0 & 0 & 1/2 \\ 0 & 0 & 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 1/2 & 1 & 0 \end{bmatrix} \begin{bmatrix} \\ \vec{x}^{(k-1)} \\ \\ \end{bmatrix} + \frac{1}{6} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \left(\alpha \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \\ \vec{x}^{(k-1)} \\ \\ \end{bmatrix} + (1-\alpha) \right)$$

Dense way: 47 FLOPs

Sparse way: 30 FLOPs

More efficient power method

- New Power method algorithm:

$$\vec{x}^{(0)} = \begin{bmatrix} 1 & 1 & \cdots & 1 \end{bmatrix}^T / n$$
$$k = 1$$
repeat
$$\vec{x}^{(k)} = \alpha P \vec{x}^{(k-1)} + \frac{\vec{e}}{n}\left(\alpha \vec{d}^T \vec{x}^{(k-1)} + (1-\alpha)\right)$$
$$\delta = \left\| \vec{x}^{(k)} - \vec{x}^{(k-1)} \right\|_\infty$$
$$k = k + 1$$
until $\delta \leq \epsilon$

Cocktail napkin computational cost analysis for Google

Operation	Approx. # FLOPs
(1) $\alpha P \vec{x}^{(k-1)}$	$14n$
(2) $\vec{d}^T \vec{x}^{(k-1)}$	n
$(1)+(2)$	n
$\vec{x}^{(k)} - \vec{x}^{(k-1)}$	n
Total	$17n = O(n)$

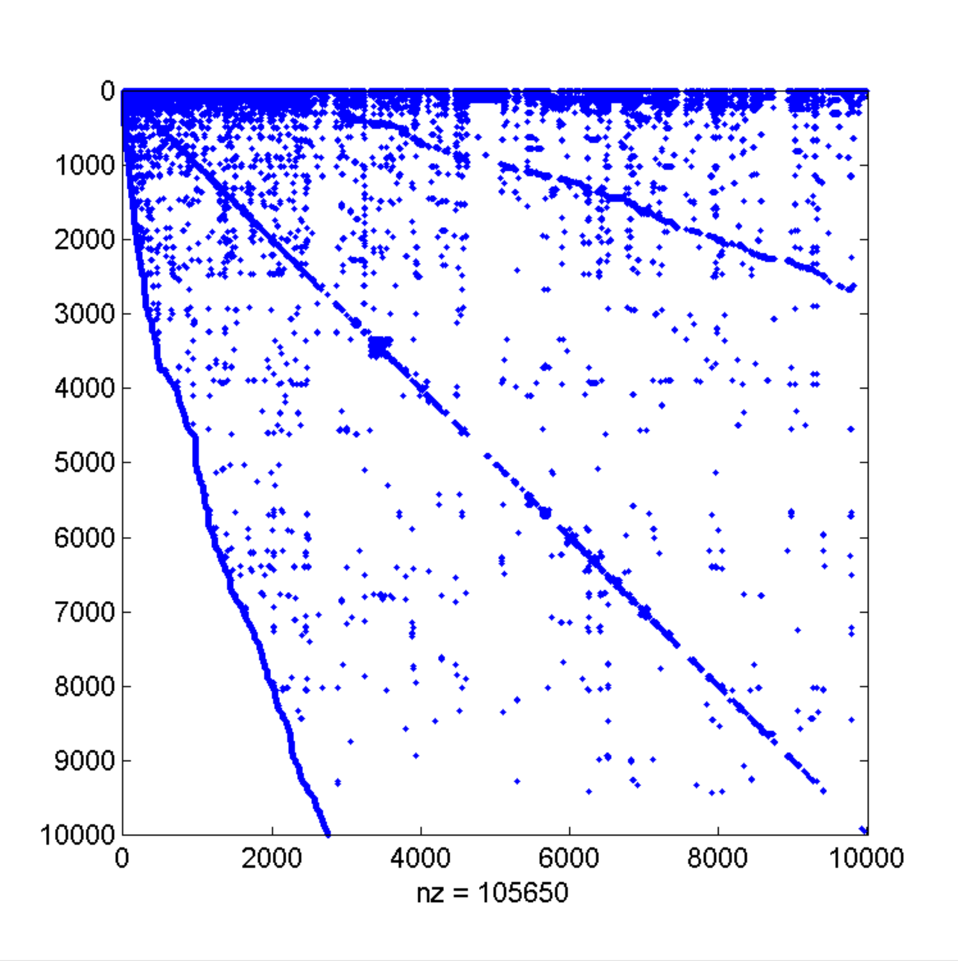
- P averages 7 nonzero entries per row.
- FLOPs $\approx 1.96 \cdot 10^{11}$
- IBM BlueGene/L: $478 \cdot 10^{12}$ FLOPs/sec
- 1 iteration: $4.09 \cdot 10^{-4}$ seconds
- 50 – 100 iterations: 0.020 – 0.041 seconds!
- In actuality it takes several days.

Example: Boise State University

- Cleve Moler's (2004) `surfer.m`
  `(http://www.mathworks.com/moler)`
- $\alpha = 0.85$
- Repeat until $\delta = \|\vec{x}^{(k)} - \vec{x}^{(k-1)}\|_\infty \leq 10^{-7}$
- Number of iterations = 44

Algorithm	Time (sec.)
Dense matrix	9.5
Sparse matrix	0.03

Connectivity matrix for www.boisestate.edu



nnz(G) = 105650
.11 % full

PageRank results for Boise State

#	PageRank	Webpage
1.	5.05933e-02	http://www.boisestate.edu
2.	1.28786e-02	http://www.boisestate.edu/maps
3.	1.24494e-02	http://www.boisestate.edu/search
4.	1.23156e-02	http://www.boisestate.edu/directory
5.	1.20138e-02	http://www.boisestate.edu/index
6.	8.55284e-03	http://broncoweb.boisestate.edu
7.	5.04204e-03	http://alumni.boisestate.edu
8.	4.10118e-03	http://blackboard.boisestate.edu
9.	4.01738e-03	http://www.boisestate.edu/community
10.	3.99552e-03	http://www.boisestate.edu/current
11.	3.98573e-03	http://www.boisestate.edu/future
12.	3.98074e-03	http://www.boisestate.edu/facultyandstaff
13.	3.62461e-03	http://library.boisestate.edu/faq/index.htm
14.	3.49803e-03	http://sspa.boisestate.edu
15.	3.00455e-03	http://admissions.boisestate.edu

PageRank results for Boise State

#	PageRank	Webpage
18.	2.66138e-03	http://coen.boisestate.edu
19.	2.57451e-03	http://selland.boisestate.edu
58.	1.20532e-03	http://www.boisestate.edu/president
107.	8.43445e-04	http://artsci.boisestate.edu
111.	8.38201e-04	http://www.boisestate.edu/biology
129.	7.53788e-04	http://math.boisestate.edu
157.	6.57107e-04	http://earth.boisestate.edu
172.	6.22599e-04	http://cobe.boisestate.edu
184.	5.90060e-04	http://cgiss.boisestate.edu
414.	3.36461e-04	http://english.boisestate.edu
472.	3.23147e-04	http://coen.boisestate.edu/cs/home.asp
562.	2.77886e-04	http://chemistry.boisestate.edu
723.	2.16570e-04	http://www.boisestate.edu/physics
2410.	7.26256e-05	http://artsci.boisestate.edu/schimpf.html
9944.	2.69934e-05	http://math.boisestate.edu/~bullock/

MATLAB code for PageRank

Editor - C:\Documents and Settings\Grady Wright\research\miscellaneous\pagerank\compute_stationary_dist_disp.m

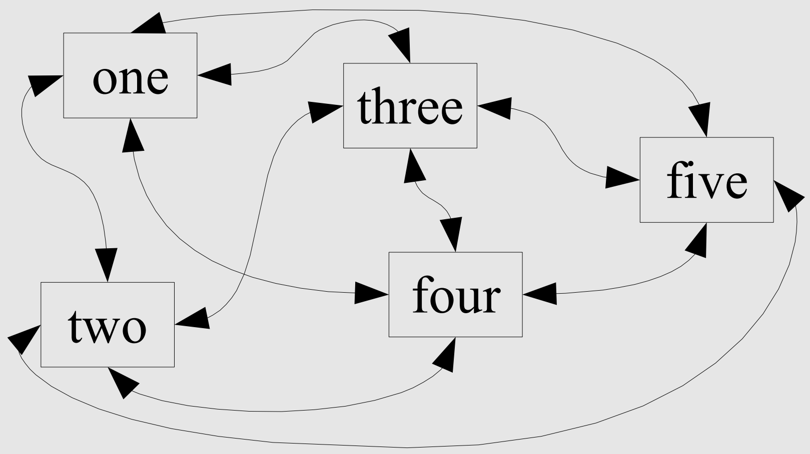
```matlab
% Collect the webpage link information:
[U,G] = surfer('http://www.boisestate.edu',2000);

n = size(G,1);
alpha = 0.85;

c = sum(G,1)/alpha;            % Compute number of out links.
id = c==0;  c(id) = 1;         % Determine the indices of the dangling nodes.
P = G*spdiags(1./c.',0,n,n);   % Compute the P transition prob. matrix.
b = double(id);                % Vector for the dangling nodes.

% Power method with sparse matrix vector products.
x = ones(n,1)/n;               % Initial guess.
y = zeros(n,1);                % temp vector.
k = 0;
while ( norm(x-y,inf) > 1e-7 & k < 800 )
   y = x;
   x = alpha*(P*y) + (alpha*(b*y) + (1-alpha))/n;
   k = k+1;
end

% Sort and print the results:
[x,id] = sort(x); id = flipud(id); x = flipud(x);
for j=1:10
   fprintf('%d  %1.2e  %s\n',j,x(j),U{id(j)});
end
```

compute_stationary_dist_disp.m | bsusurfer.m

script | Ln 26 | Col 4 | OVR

Exercises

1. What is the PageRank vector for the following web:



2. What effect does decreasing α have on the PageRank model?

Recall: $$A = \alpha P + \alpha \frac{\vec{e}}{n} \vec{d}^T + (1-\alpha) \frac{\vec{e}}{n} \vec{e}^T, \quad 0 < \alpha < 1$$

3. Let the web W consist of n pages and suppose $\vec{x}$ satisfies $\sum_{j=1}^{n} x_j = 1$.

   If $\vec{y} = A\vec{x}$, show $\sum_{j=1}^{n} y_j = 1$, where A is given above.

More advanced PageRank topics

1. Effect of changing the teleport probability α.

$$\vec{x}^{(k+1)} = \alpha P \vec{x}^{(k)} + \frac{\vec{e}}{n}\left(\alpha \vec{d}^{T} \vec{x}^{(k)} + (1-\alpha)\right) = \vec{v} + O(\alpha^{k})$$

2. Faster algorithms.

$$A\vec{v} = \vec{v} \qquad \Leftrightarrow \qquad \left[I - \alpha\left(P + \frac{\vec{e}}{n}\vec{d}\right)\right]\vec{v} = \frac{(1-\alpha)}{n}\vec{e}$$

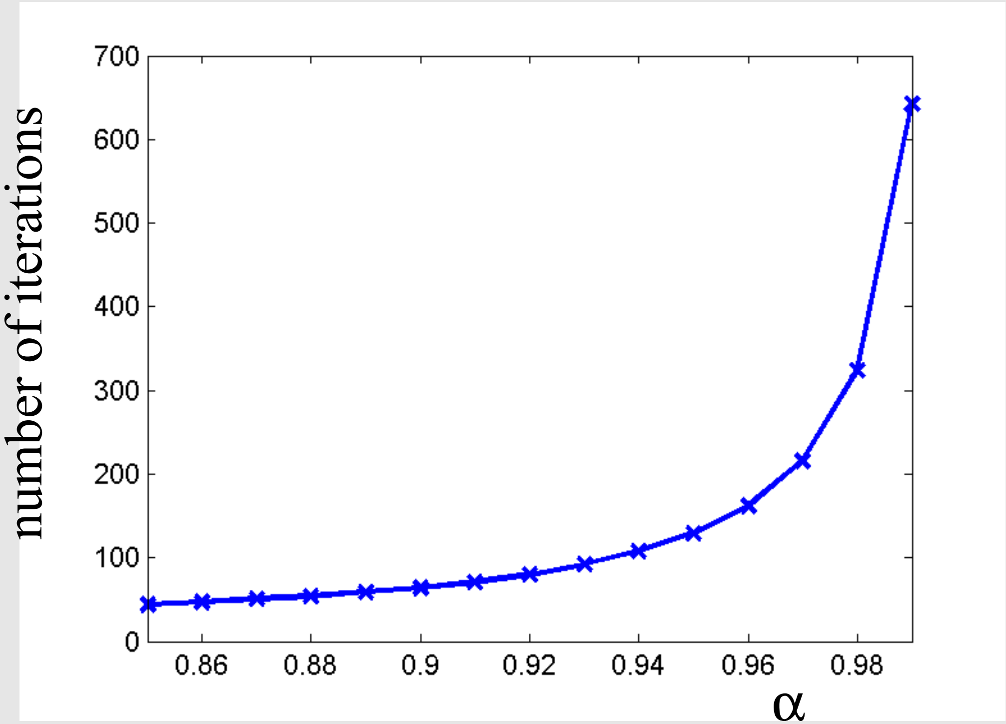
3. Personalizing PageRank.

$$A = \alpha P + \alpha \frac{\vec{e}}{n}\vec{d}^{T} + (1-\alpha)\frac{\vec{e}}{n}\vec{e}^{T} \qquad \Rightarrow \qquad A_u = \alpha P + \alpha \vec{u}\,\vec{d}^{T} + (1-\alpha)\vec{u}\,\vec{e}^{T}$$

4. Search engine optimization (SEO).

5. Updating the PageRank vector.

Effect of α

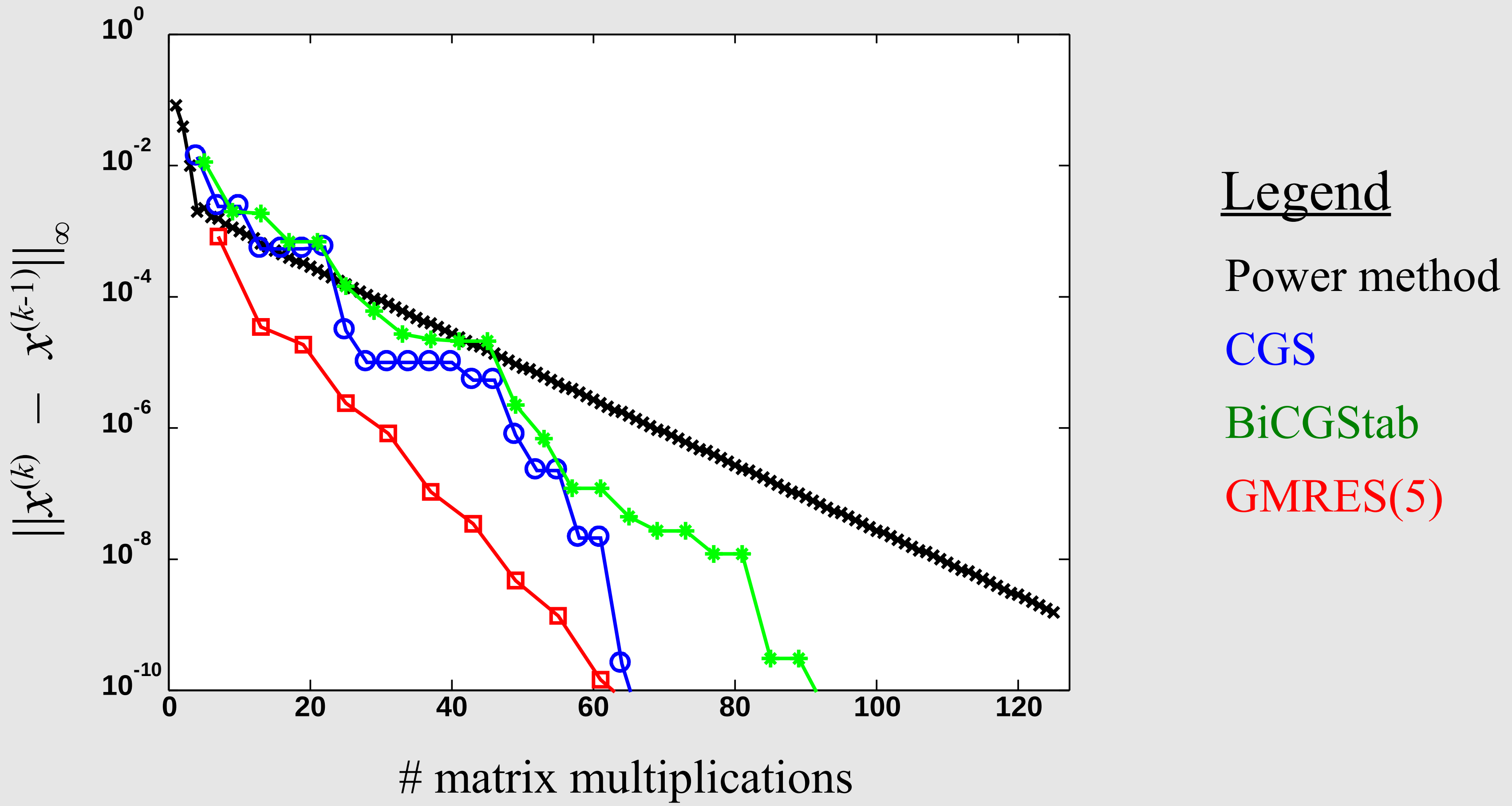


- Repeat until $\delta = \|\vec{x}^{(k)} - \vec{x}^{(k-1)}\|_\infty \leq 10^{-7}$
- $\vec{x}^{(k+1)} = \vec{v} + O(\alpha^k)$
- # iterations $\approx \dfrac{\log \delta}{\log \alpha}$

PageRank	α=0.85	α=0.95
6.	http://broncoweb.boisestate.edu	http://broncoweb.boisestate.edu
7.	http://alumni.boisestate.edu	http://alumni.boisestate.edu
8.	http://blackboard.boisestate.edu	http://www.boisestate.edu/community
9.	http://www.boisestate.edu/community	http://www.boisestate.edu/current
10.	http://www.boisestate.edu/current	http://www.boisestate.edu/future
11.	http://www.boisestate.edu/future	http://www.boisestate.edu/facultyandstaff
12.	http://www.boisestate.edu/facultyandstaff	http://blackboard.boisestate.edu
13.	http://library.boisestate.edu/faq/index.htm	http://rec.boisestate.edu/beatpete/index.cfm
14.	http://sspa.boisestate.edu	http://rec.boisestate.edu/beatpete/contact.cfm
15.	http://admissions.boisestate.edu	http://rec.boisestate.edu/beatpete/sponsors.cfm

Faster algorithms

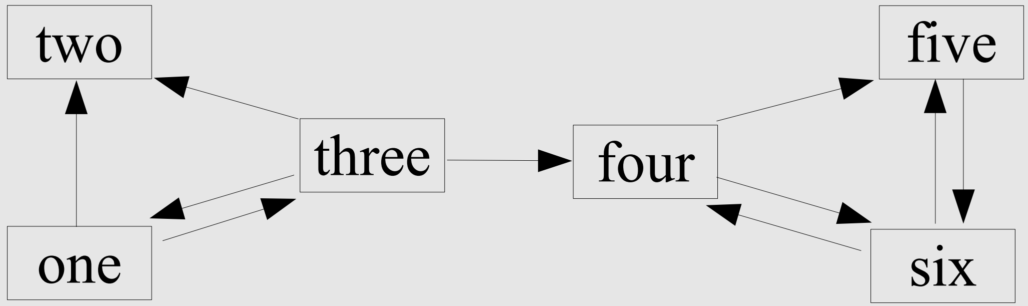
- Idea: $A\vec{v}=\vec{v} \quad \Leftrightarrow \quad \left[I-\alpha\left(P+\frac{\vec{e}}{n}\vec{d}\right)\right]\vec{v}=\frac{(1-\alpha)}{n}\vec{e}$
- Can use advanced iterative methods for sparse linear systems!



Personalizing PageRank

$$A = \alpha P + \alpha \frac{\vec{e}}{n}\vec{d}^T + (1-\alpha)\frac{\vec{e}}{n}\vec{e}^T \quad \Rightarrow \quad A_u = \alpha P + \alpha \vec{u}\,\vec{d}^T + (1-\alpha)\vec{u}\,\vec{e}^T$$

- Example:

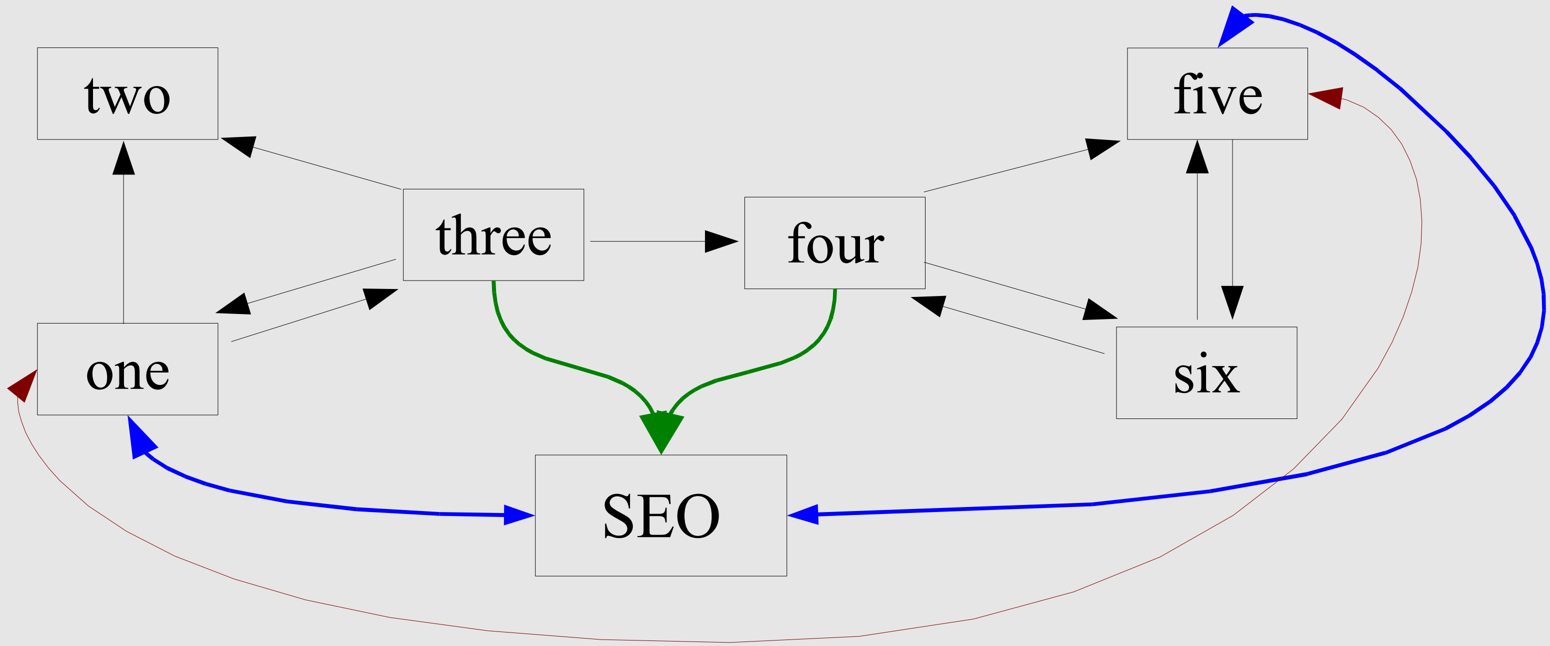


Personalization vector: $\vec{u} = \begin{bmatrix} 1/4 & 1/8 & 1/4 & 1/4 & 1/16 & 1/16 \end{bmatrix}^T$

$$A = \alpha \begin{bmatrix} 0 & 1/4 & 1/3 & 0 & 0 & 0 \\ 1/2 & 1/8 & 1/3 & 0 & 0 & 0 \\ 1/2 & 1/4 & 0 & 0 & 0 & 0 \\ 0 & 1/4 & 1/3 & 0 & 0 & 1/2 \\ 0 & 1/16 & 0 & 1/2 & 0 & 1/2 \\ 0 & 1/16 & 0 & 1/2 & 1 & 0 \end{bmatrix} + (1-\alpha) \begin{bmatrix} 1/4 & 1/4 & 1/4 & 1/4 & 1/4 & 1/4 \\ 1/8 & 1/8 & 1/8 & 1/8 & 1/8 & 1/8 \\ 1/4 & 1/4 & 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 & 1/4 & 1/4 \\ 1/16 & 1/16 & 1/16 & 1/16 & 1/16 & 1/16 \\ 1/16 & 1/16 & 1/16 & 1/16 & 1/16 & 1/16 \end{bmatrix}$$

Search engine optimization

- Idea



<u>Before SEO</u> → <u>After SEO</u>

$$\vec{v} = \begin{bmatrix} 0.051704746 \\ 0.073679263 \\ 0.057412413 \\ 0.19990381 \\ 0.26859608 \\ 0.34870368 \end{bmatrix} \rightarrow \begin{bmatrix} 0.12742415 \\ 0.061268076 \\ 0.050530372 \\ 0.11365423 \\ 0.26626095 \\ 0.17423120 \\ 0.20663101 \end{bmatrix}$$

Concluding remarks

- PageRank is the “Heart of Google software”
- Use random walk (surf) to formulate PageRank problem.
- Use linear algebra to define PageRank.
- Use the power method to compute PageRank.
- Tools for PageRank are readily available and accessible to undergraduates!
- Many new areas of PageRank research to explore.