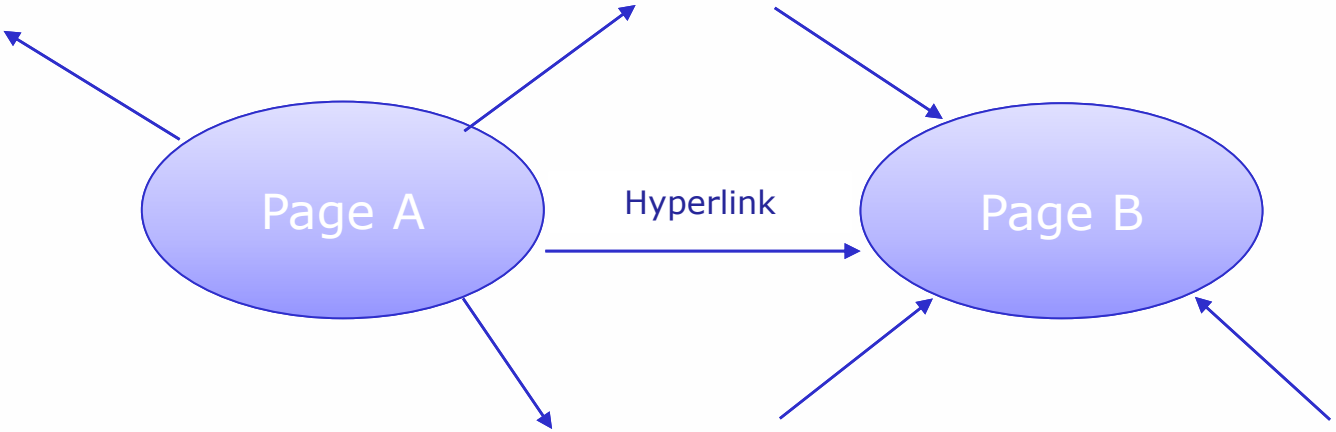
POLITECNICO DI MILANO

Link Analysis

Information Retrieval and Data Mining

Prof. Matteo Matteucci

Hyperlinks for Indexing and Ranking



Intuitions

- The ***anchor text*** might describe the target page B
  - Anchor text indexing
- The ***hyperlink*** from A to B connotes a conferral of authority on page B, by the creator of page A
  - Link based ranking

Anchor Text Indexing (1)

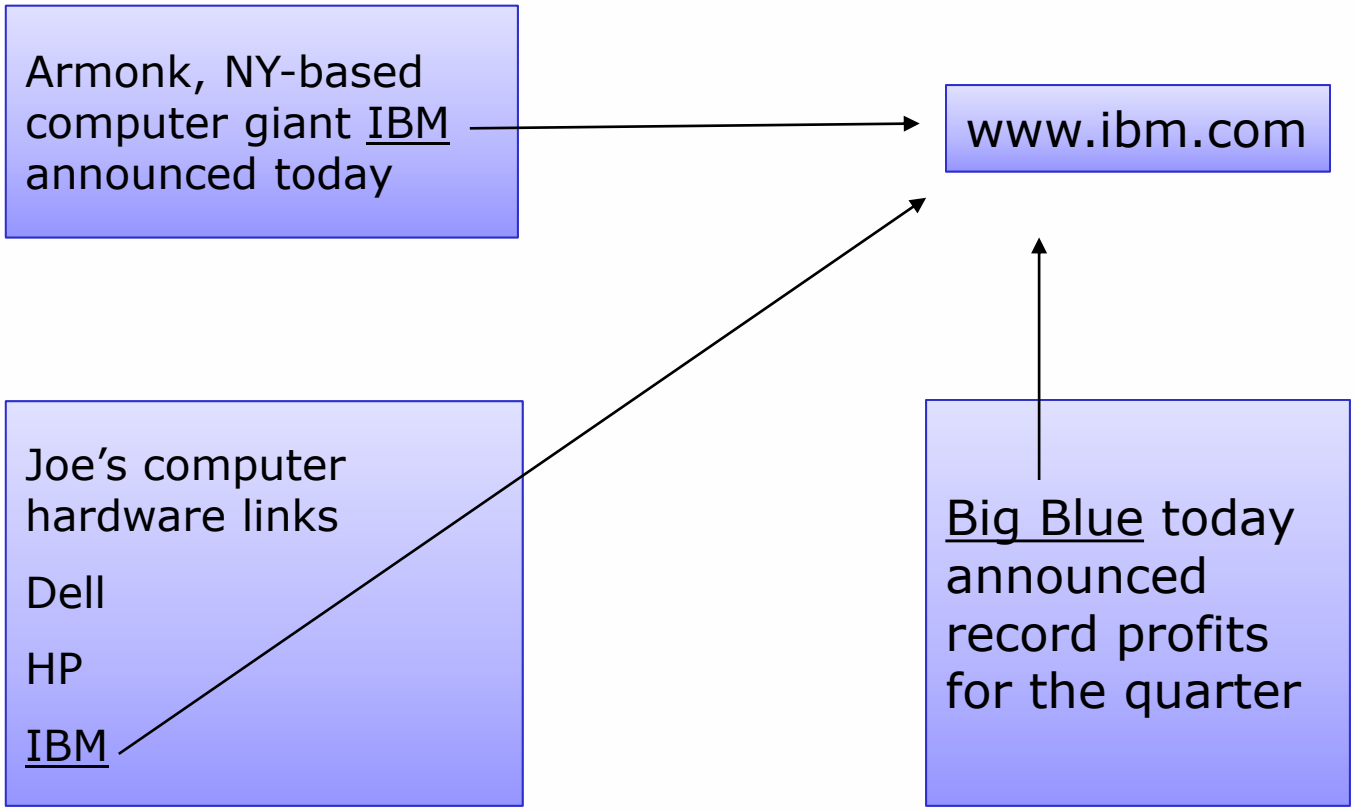
Issues of conventional inverted index search:

- Sometims page B does not provide a description of itself
  - http://www.ibm.com page not contain “computer”
- Gap between how a page presents itself and how web users would describe it (e.g., Big Blue -> IBM)
- Many pages embed text in graphics and images, making HTML parsing inefficient

Solution:

- Include anchor text terms in inverted indexing
- Weight anchor text terms based on frequency (to penalize words such as “click” or “here”)

Anchor Text Indexing (2)



Link Based Ranking

- Second generation search engine identify relevance
  - topic-relatedness (boolean, vector, probabilistic, …)
  - authoritativeness (link base ranking)

The Web can be seen as a network of recommendations

- Link analysis / centrality indices used for 60+ years
  - Sociology
  - Psychology
  - Bibliometrics
  - Information Retrieval
  - …

Why Link Analysis?

The WWW can been seen as a (directed) graph:

- Vertices: web pages
- Edges: hyperlinks

Goal: use of hyperlinks to index/rank Web search results

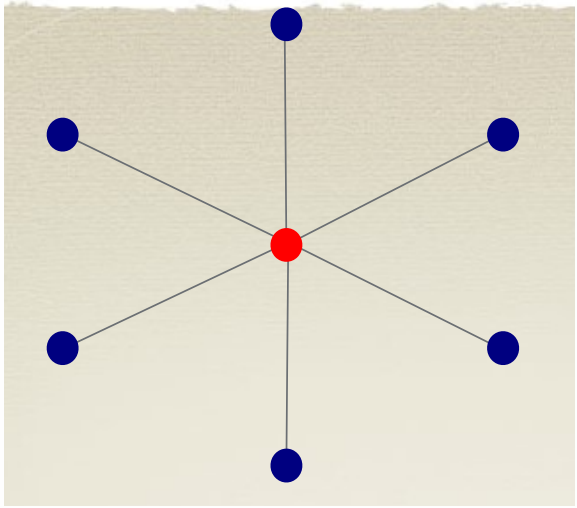
The same can be done for other interlinked environments:

- Dictionaries
- Encyclopedias
- Scientific publications
- Social networks
- …

Centrality Indeces

- Several centrality indices exists:
  - *Spectral indices*, based on linear algebra construction
  - *Path-based indices*, based on the number of paths or shortest paths (geodesics) passing through a vertex
  - *Geometric indices*, based on distances from a vertex to other vertices

- The center of a star is more important than the other nodes
  - It has largest degree
  - It is on the shortest paths
  - It is maximally close to everybody



(In-)Degree Centrality (1)

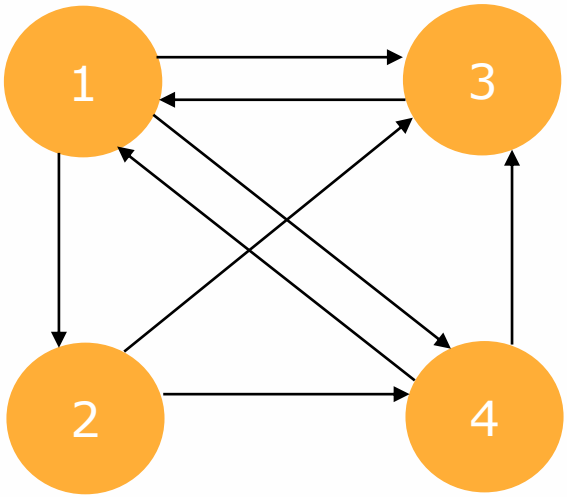
Basics assumptions

- Any Web page is assigned a non-negative score
- A page score is derived from links made to that page from other web pages
- Links to a given page are called backlinks
- Web democracy: pages vote for the importance of other pages by linking to them

(In-)degree centrality

- Count the number of incoming links $x_k = \sum_{i \to k} 1$
- (Or the nodes at distance 1)

(In-)Degree Centrality (2)



$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \\ 3 \\ 2 \end{bmatrix}$$

An important issue:

- Ignores the fact that a link to page ***j*** from an important page should boost page ***j***'s importance score more than a link from an unimportant one
- Indeed a link to your homepage from ***www.yahoo.com*** should boost your page's score more than a link from ***www.foobar.com***

Betweenness Centrality (1971)

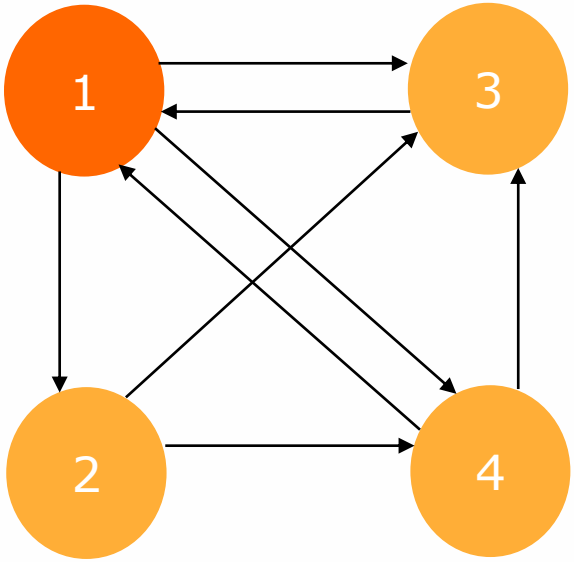
- Count the fraction of ***shortest paths*** from *i* to *j* passing through node *k*

$$x_k = \sum_{i,j \neq k} \frac{\sigma_{ij}(k)}{\sigma_{ij}}$$

- Often scaled dividing by *(n-1)/(n-2)* in directed graphs

- Let Consider *Node 1*
  - $2 \rightarrow 3$
  - $2 \rightarrow 4$
  - $3 \rightarrow 2$
  - $3 \rightarrow 4$
  - $4 \rightarrow 2$
  - $4 \rightarrow 3$

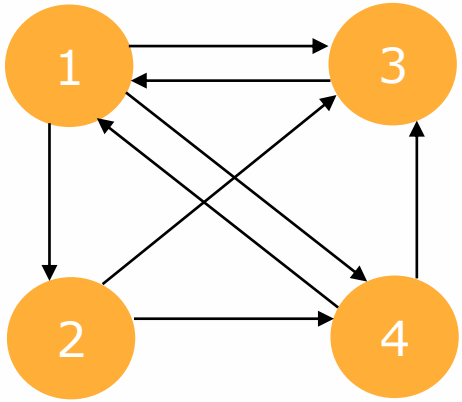
$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 3/6 \\ 0 \\ 0 \\ 1/6 \end{bmatrix}$$



Kats Centrality (1953)

- Count the number of paths of length *t* ending at node *k*
- Discount this number by α^t

$$x_k = \sum_{t=0}^{+\infty} \alpha^t \Pi_k(t) = \sum_{t=0}^{+\infty} \sum_{j=1}^{n} \alpha^t (E^t)_{jk}$$



- Lets Consider *Node 1*

$$x_1 = \alpha \sum_{j=1}^{n} (E)_{j1} + \alpha^2 \sum_{j=1}^{n} (E^2)_{j1} + \dots$$
$$= 2\alpha + 5\alpha^2 + \dots$$

Number of paths of length 1 from node j to node 1

$$\mathbf{E} = \begin{bmatrix} 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \end{bmatrix}$$

Number of paths of length 2 from node j to node 1

$$\mathbf{E}^2 = \begin{bmatrix} 2 & 0 & 2 & 1 \\ 2 & 0 & 1 & 0 \\ 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix}$$

- Computed efficiently by

$$\mathbf{x} = ((\mathbf{I} - \alpha \mathbf{E}^T)^{-1} - \mathbf{I})\mathbf{1}$$

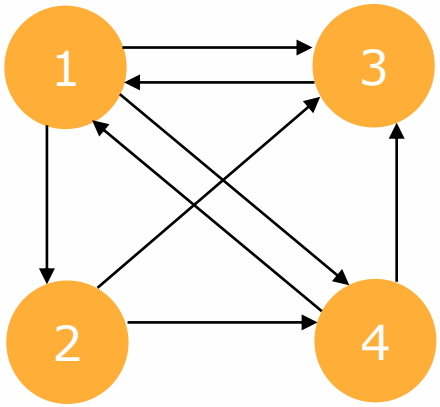
Column vector of *n* ones

Closeness Centrality (1950)

- Geometric-based: inverse of the sum of the distances from each node

$$x_k = \frac{1}{\sum_{j \neq k} d(j,k)}$$

- The summation is over all nodes such that $d(j,k) < \infty$
- Lets consider *Node 1*



$$x_1 = \frac{1}{2+1+1} = \frac{1}{4}$$

$$x_2 = \frac{1}{1+2+2} = \frac{1}{5}$$

$$x_3 = \frac{1}{1+1+1} = \frac{1}{3}$$

$$x_4 = \frac{1}{1+1+2} = \frac{1}{4}$$

- **Harmonic Centrality** (2000) includes infinity as well ...

Wei Index (1953)

- Spectral-based: if a matrix represents whether a team defeated another team, we can define a general score by iteratively computing the sum of the scores of the teams that have been defeated

- In our context:
  - If page *j* links to page *k*, we boost page *k*'s score by x_j

$$x_k = \sum_{j \in L_k} x_j$$

L_k is the set of page k's backlinks

  - *Issue:* a single page gains influence just by linking lots of other pages

Seeley Index (1949)

- Spectral-based: in a group of children, a child is as popular as the sum of the popularities of the children who like him, but popularities are divided evenly among friends

- In our context:
  - If page j contains n_j links, one of which to page k, we will boost page k's score by x_j/n_j instead than by x_j

$$x_k = \sum_{j \in L_k} \frac{x_j}{n_j}$$

L_k is the set of page k's backlinks

n_j is the number of page j's outgoing links

  - Each page gets one vote, which is divided up among its outgoing links

Seeley Index (1949)

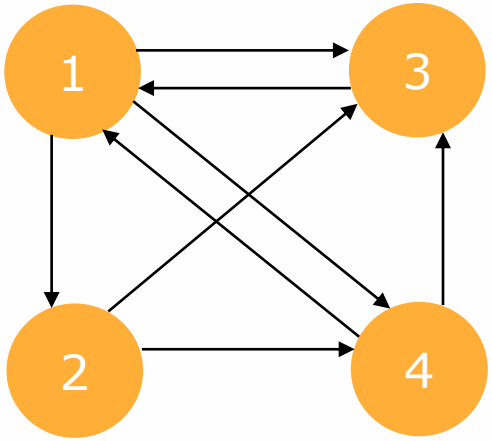
$$x_1 = \frac{x_3}{1} + \frac{x_4}{2}$$

$$x_2 = \frac{x_1}{3}$$

$$x_3 = \frac{x_1}{3} + \frac{x_2}{2} + \frac{x_4}{2}$$

$$x_4 = \frac{x_1}{3} + \frac{x_2}{2}$$

$$\mathbf{Ax} = \mathbf{x}$$



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

$\mathbf{A}_{ij} = 0$ if there is no link between page *j* and *i*

$\mathbf{A}_{ij} = \frac{1}{n_j}$ otherwise, with n_j the number of outgoing links of page *j*

- Solve the eigensystem for the eigenvector corresponding to eigenvalue 1
- The matrix A is the transpose of the (weighted) adjacency matrix of the Web graph

Seeley Index (1949)

- Can be obtained by the power iteration method

$$\mathbf{x} = \lim_{k \to \infty} \mathbf{A}^k \mathbf{x}_0$$

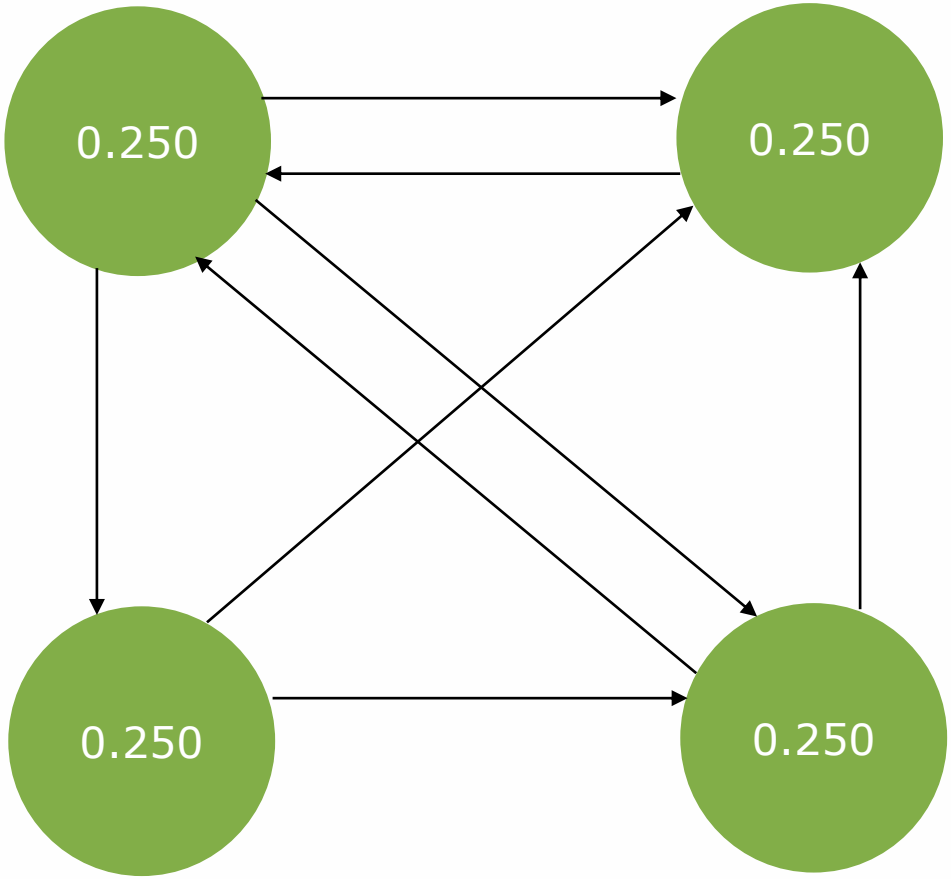
Steady state in Ergothic Markov Chain!!!

  - where $\mathbf{x}_0$ is some initial column vector with non-zero entries
  - Convergence is guaranteed for strongly connected graphs (i.e. if you can get from any page to any other page in a finite number of steps)

- A surfer moves from one page to next randomly choosing one of the outgoing links
- The component x_j of the normalized score vector $\mathbf{x}$ is the time the surfer spends, in the long run, on page j of the web
- More important pages tend to be linked to by many other pages and so the surfer hits those most often.

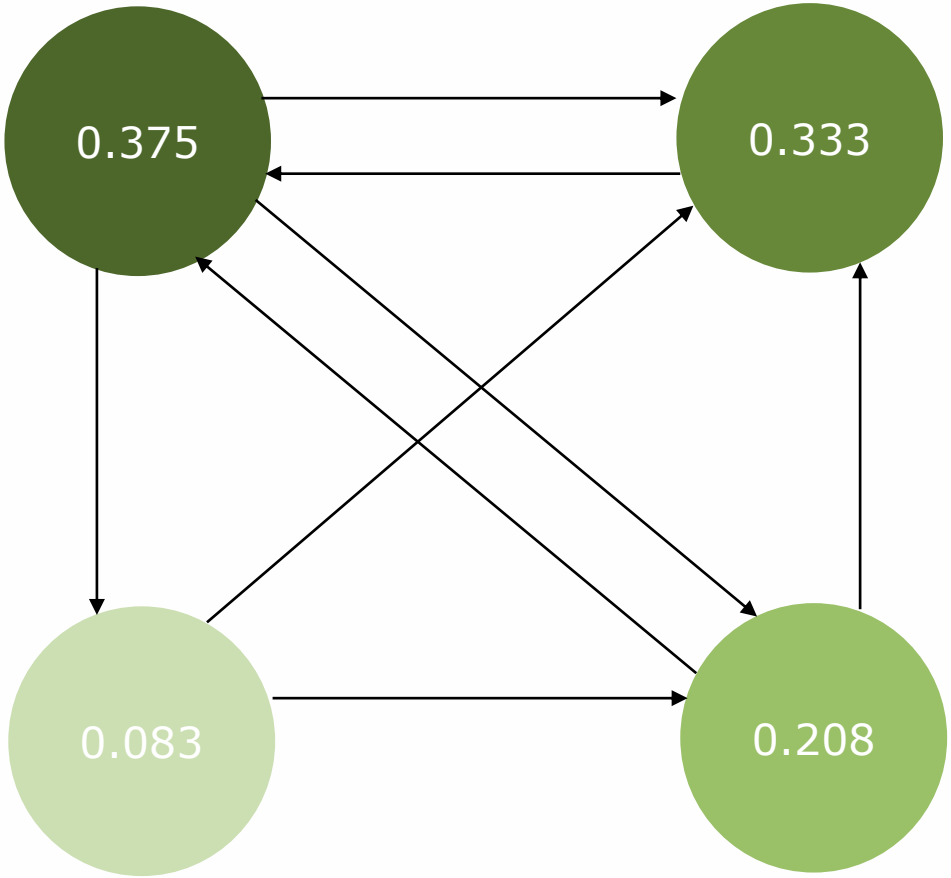
Seeley Example

- Power iteration k=0



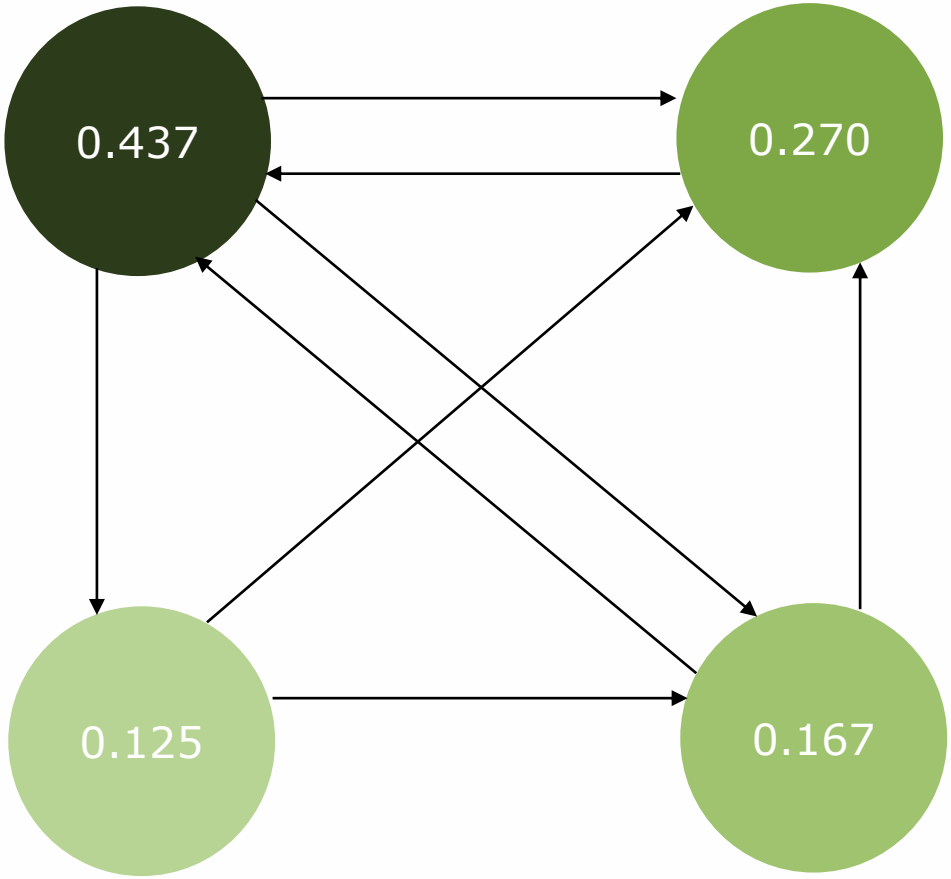
Seeley Example

- Power iteration k=1



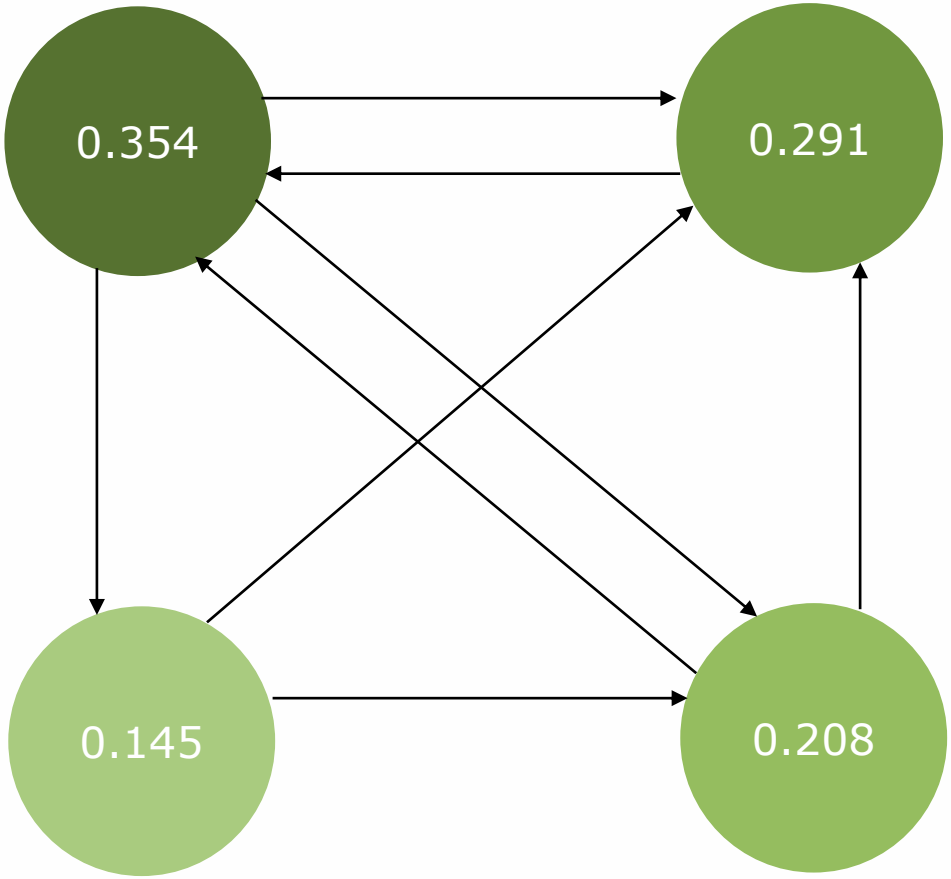
Seeley Example

- Power iteration k=2



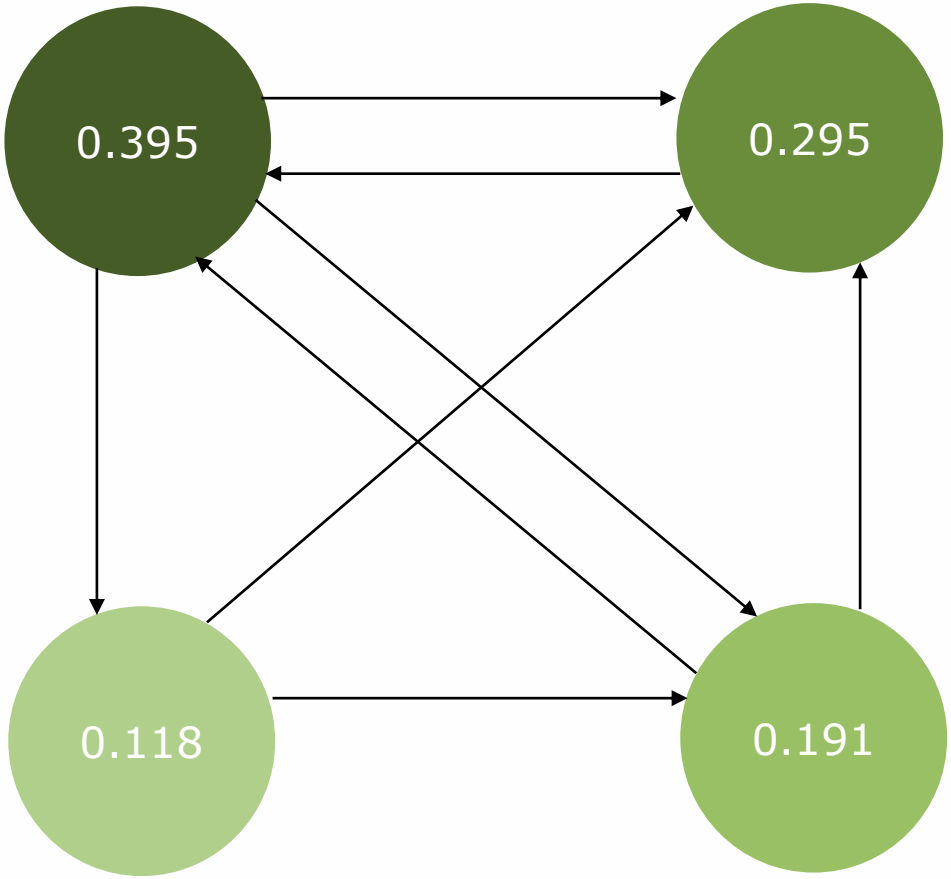
Seeley Example

- Power iteration k=3



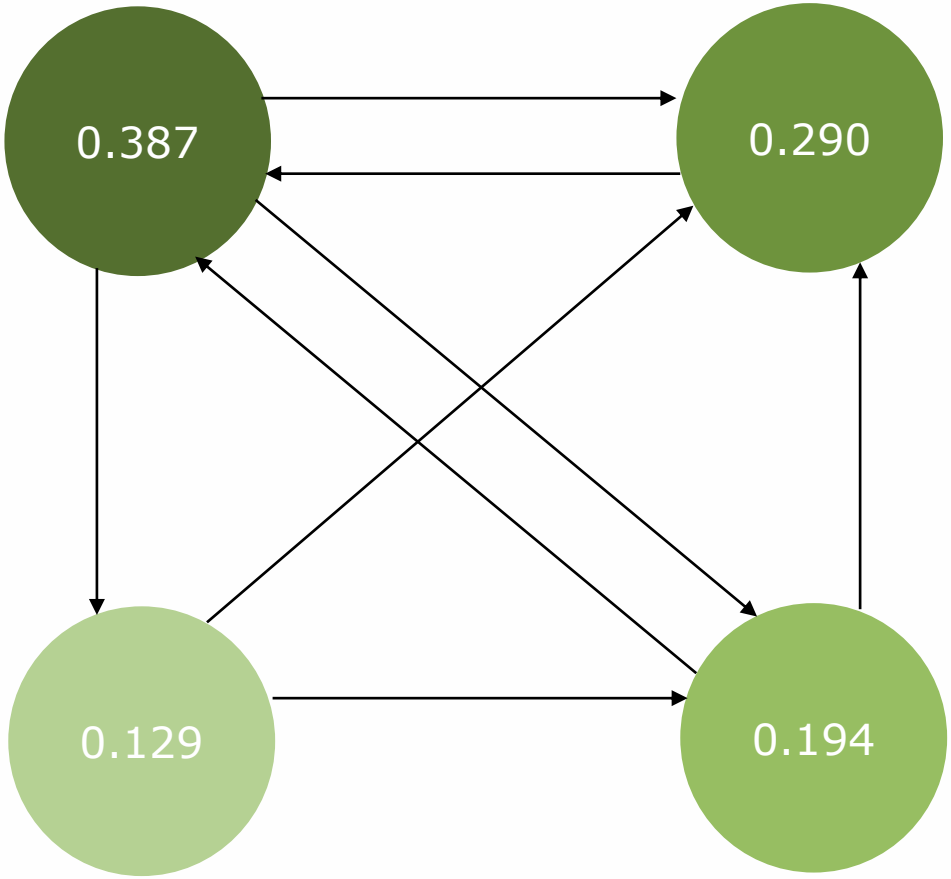
Seeley Example

- Power iteration k=4



Seeley Example

- Power iteration $k \rightarrow \infty$



PageRank (1999)

- Spectral-based: start from Seeley's equation and add an adjustment to make it have a unique solution

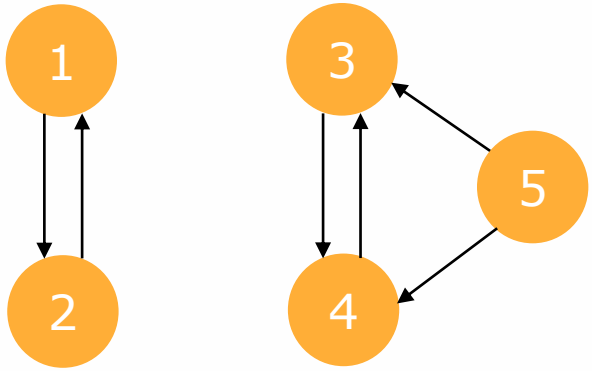
$$x_k = (1-m)\sum_{j \in L_k} \frac{x_j}{n_j} + m\frac{1}{n}$$

- Enhanced Random surfer moves from one page to next
  - If the surfer is currently at a page with r outgoing links,
    - with probability 1-m he randomly chooses one of these links
    - with probability m he jumps to any randomly selected page
- Equivalent to adding implicit links from each page to any other page; the resulting graph is strongly connected

Uniqueness of Ranking (1)

- Power iterations may not always converge to a unique ranking when the graph is not strongly connected

- Example:



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

- Equation x = Ax is satisfied by any linear combination of:

$$\mathbf{x}^{(1)} = [1/2 \quad 1/2 \quad 0 \quad 0 \quad 0]^T$$

$$\mathbf{x}^{(2)} = [0 \quad 0 \quad 1/2 \quad 1/2 \quad 0]^T$$

Uniqueness of Ranking (2)

- To fix this, at each step the random surfer can randomly jump to any page of the Web with some probability m being (1 – m) a damping factor
- We can generate unique importance scores by modifying the matrix A as follows

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

where $\boldsymbol{S} \in \mathbb{R}^{nxn}, S_{i,j} = 1/n$

$$0 \leq m < 1$$

- The PageRank score is computed as

$$\mathbf{x} = \lim_{k \to \infty} \mathbf{M}^k \mathbf{x}_0$$

Uniqueness of Ranking (3)

- For example, setting m = 0.15

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1/2 \\ 0 & 0 & 1 & 0 & 1/2 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \Rightarrow \mathbf{M} = \begin{bmatrix} 0.03 & 0.88 & 0.03 & 0.03 & 0.03 \\ 0.88 & 0.03 & 0.03 & 0.03 & 0.03 \\ 0.03 & 0.03 & 0.03 & 0.88 & 0.455 \\ 0.03 & 0.03 & 0.88 & 0.03 & 0.455 \\ 0.03 & 0.03 & 0.03 & 0.03 & 0.03 \end{bmatrix}$$

- The unique PageRank score is given by

$$\mathbf{x} = \begin{bmatrix} 0.2 & 0.2 & 0.285 & 0.285 & 0.03 \end{bmatrix}^T$$

- Instead of

$$\mathbf{x}^{(1)} = \begin{bmatrix} 1/2 & 1/2 & 0 & 0 & 0 \end{bmatrix}^T$$

$$\mathbf{x}^{(2)} = \begin{bmatrix} 0 & 0 & 1/2 & 1/2 & 0 \end{bmatrix}^T$$

Dealing with Dangling Nodes (1)

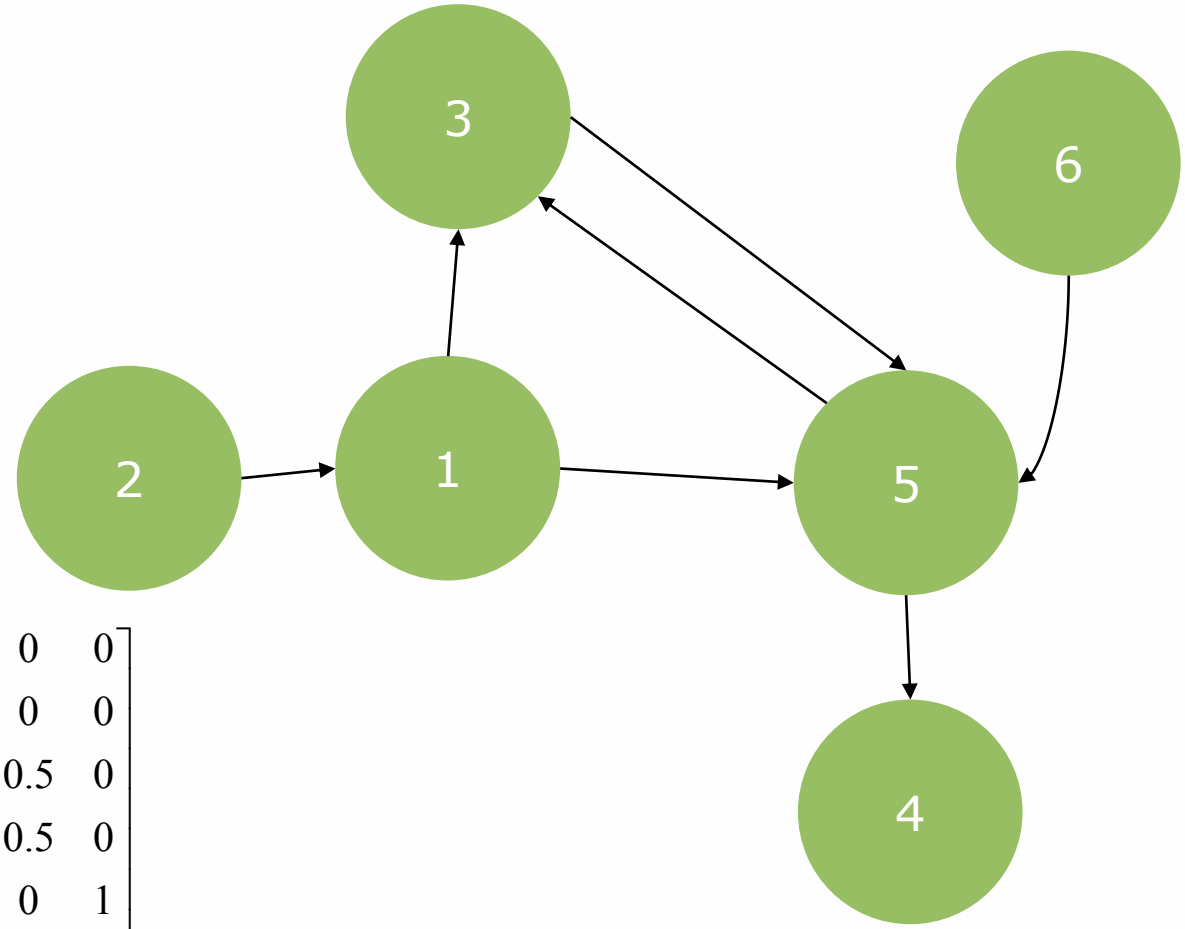
- Nodes without outgoing links, e. g., webpages with no links to other pages:
  - image & music files
  - pdf files
  - pages whose links haven't been crawled
  - protected pages
- Some estimates say more than 50%, others say between 60% and 80%
- A web with dangling nodes produces a matrix A which contains one or more columns of all zeros. Matrix A does not have an eigenvalue equal to 1

Dealing with Dangling Nodes (2)

- Solution 1:
  - Remove dangling nodes
  - build matrix $\mathbf{M}$, with $m > 0$
  - Compute the importance scores for the remaining nodes
  - Reinsert dangling nodes, computing their importance scores based on incoming links only
- Solution 2
  - replace the all-zero columns of matrix **A** corresponding to dangling nodes with $1/n$ term
  - equivalent to browsing to a new page at random when the user reaches a page with no outgoing links
  - build matrix $\mathbf{M}$, with $m > 0$
  - Compute the importance scores

PageRank Example (Solution 1)

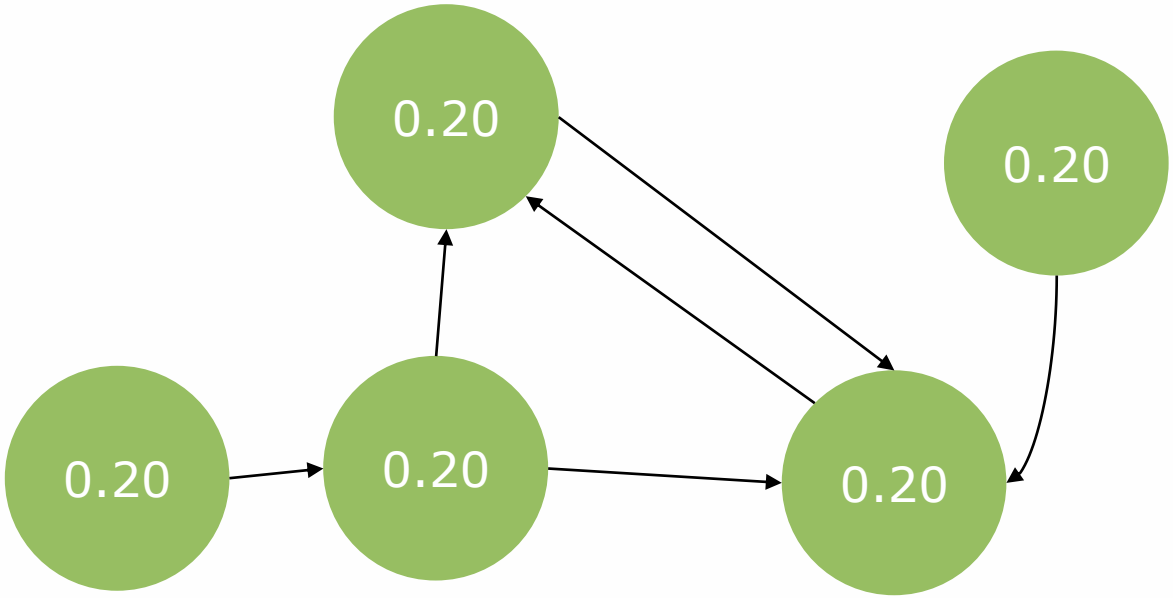
- Remove dangling nodes



$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0.5 & 0 & 0 & 0 & 0.5 & 0 \\ 0 & 0 & 0 & 0 & 0.5 & 0 \\ 0.5 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

PageRank Example (Solution 1)

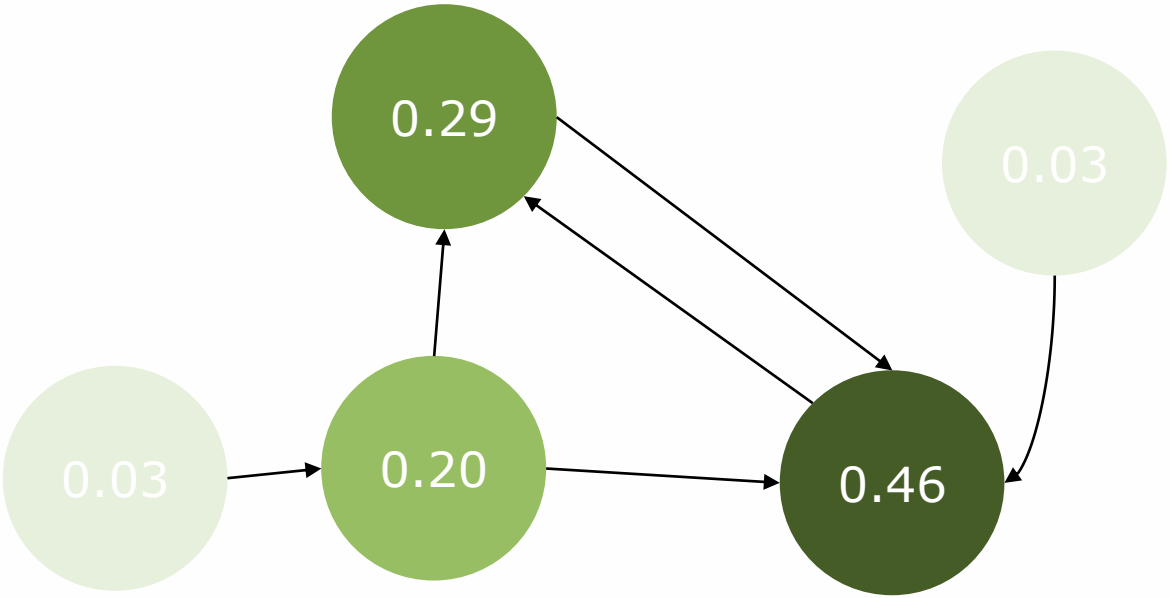
- Power iteration k = 0



$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0.5 & 0 & 0 & 0 & 0.5 & 0 \\ 0 & 0 & 0 & 0 & 0.5 & 0 \\ 0.5 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \Rightarrow \mathbf{A} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0.5 & 0 & 0 & 1 & 0 \\ 0.5 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

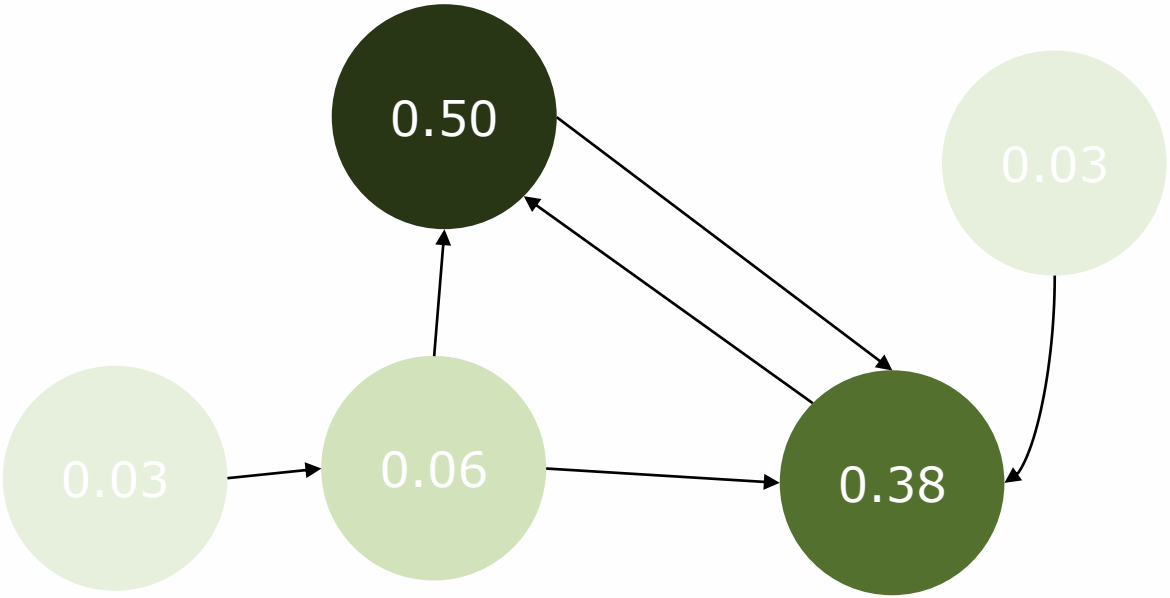
PageRank Example (Solution 1)

- Power iteration k = 1 (with dumping m = 0.15)



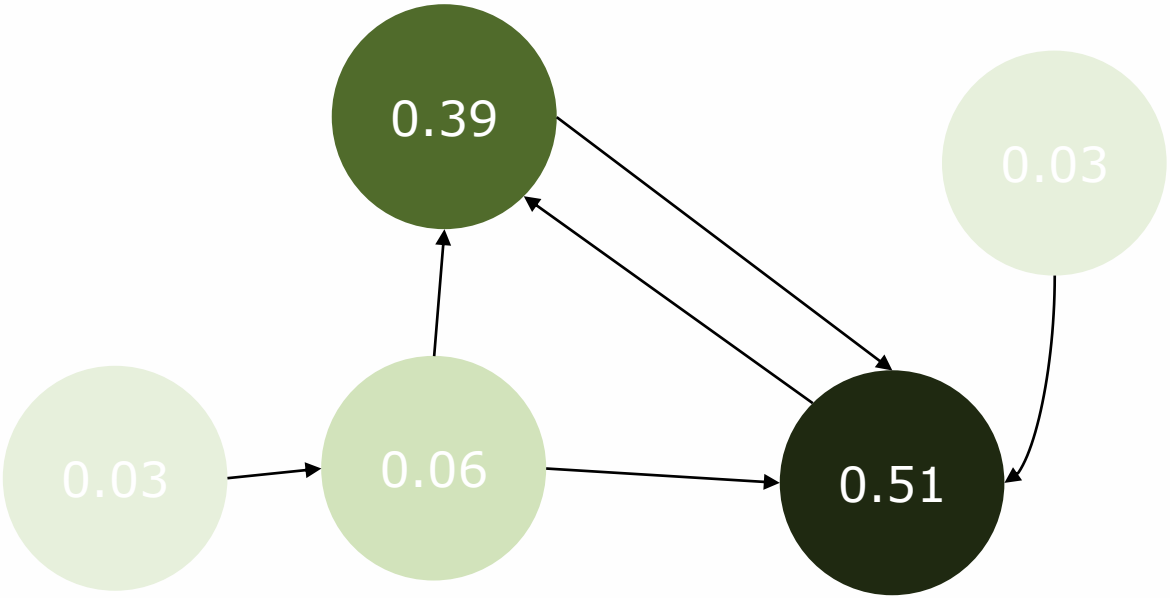
PageRank Example (Solution 1)

- Power iteration k = 2 (with dumping m=0.15)



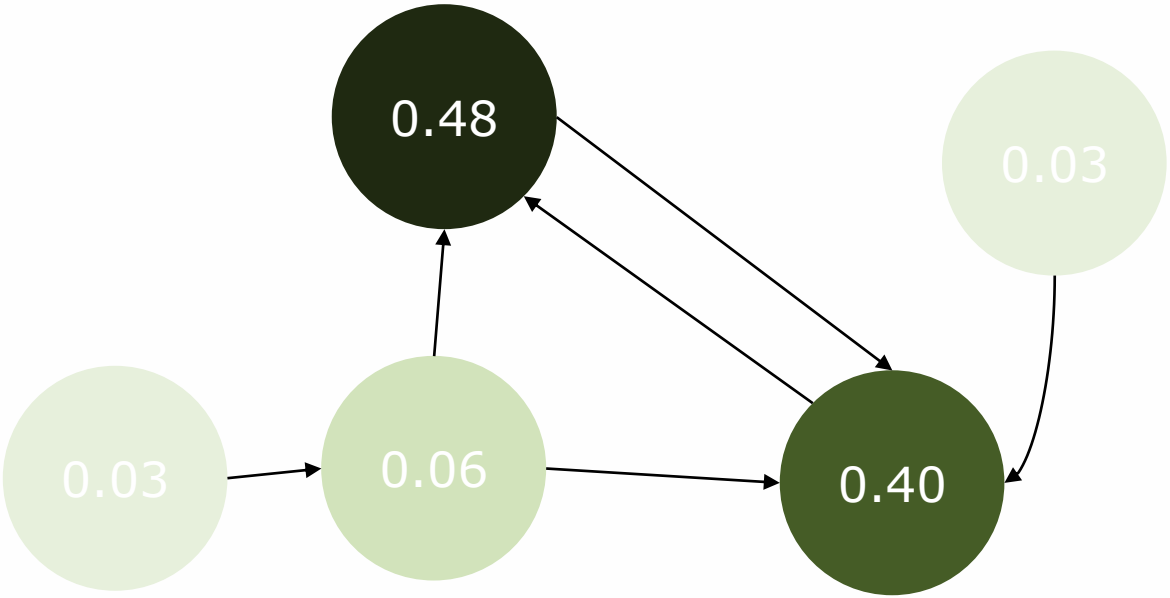
PageRank Example (Solution 1)

- Power iteration k = 3 (with dumping m=0.15)



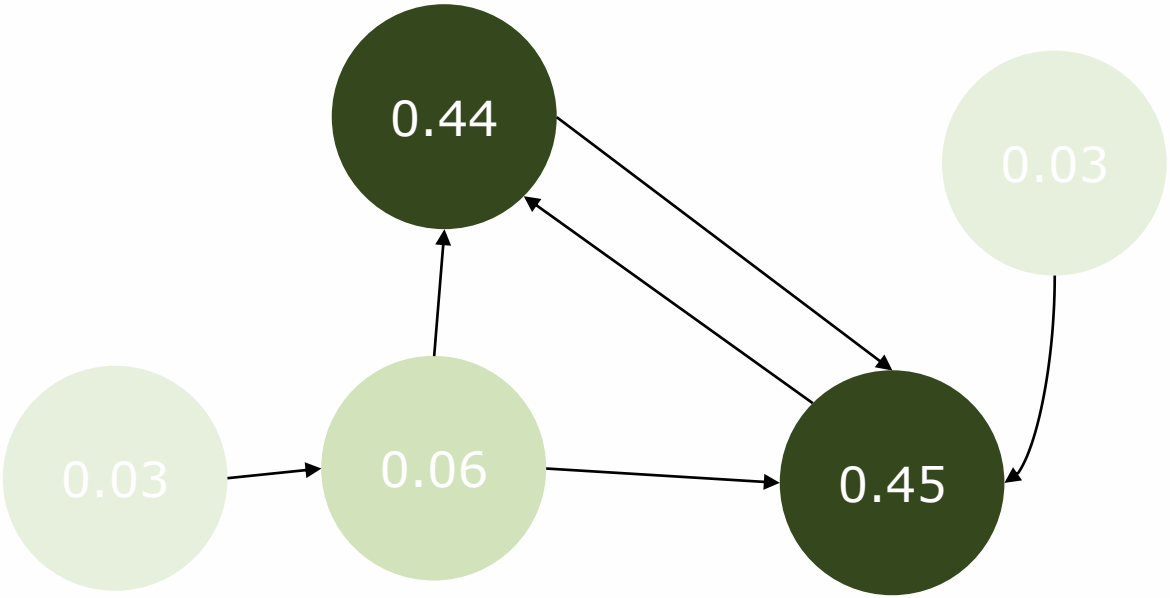
PageRank Example (Solution 1)

- Power iteration k = 4 (with dumping m=0.15)



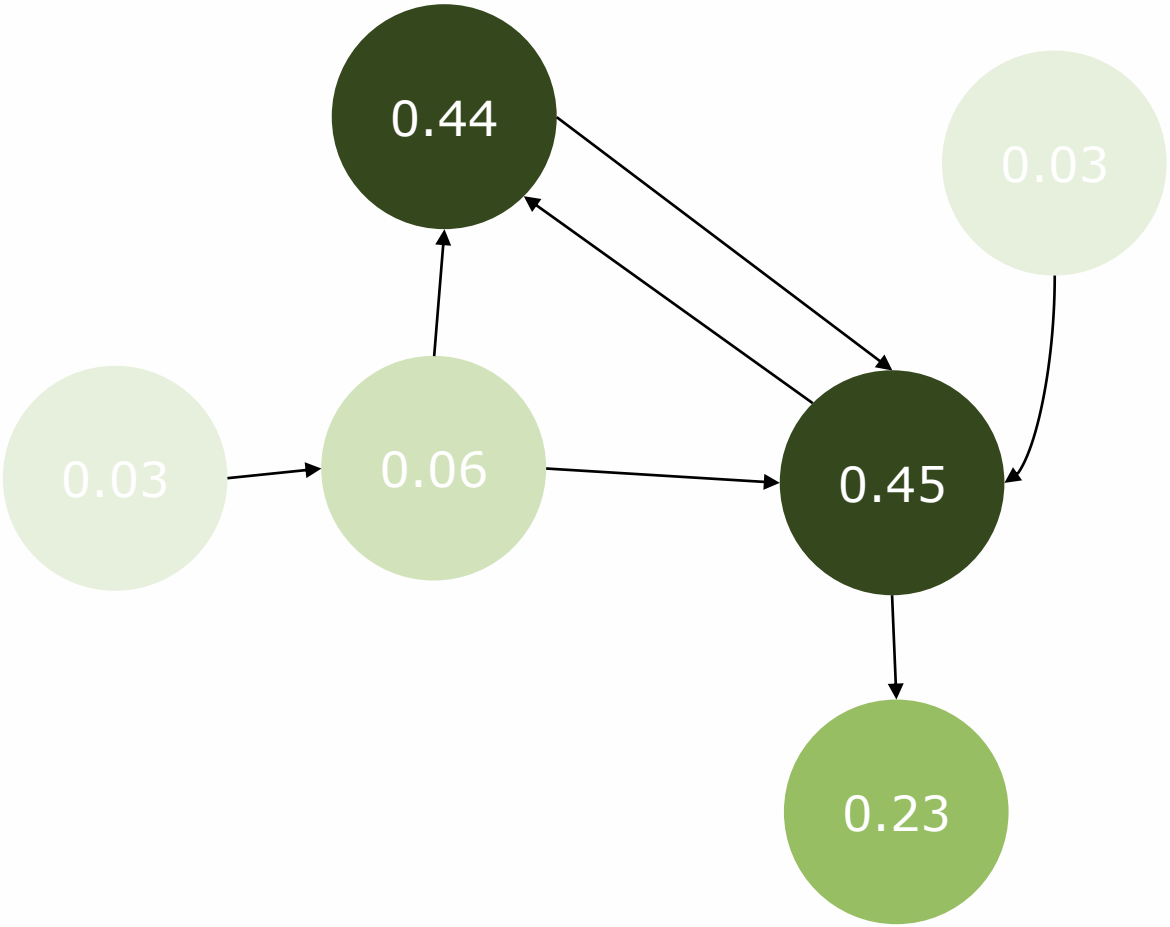
PageRank Example (Solution 1)

- Power iteration $k \to \infty$ (with dumping m=0.15)



PageRank Example (Solution 1)

- Add dangling node (x4 = x5 / 2)



References (some)

- S. Chakrabarti, "Mining the Web", *Morgan Kaufman*, 2003
- J. Kleinberg, "Authoritative sources in a hyperlinked environment", *Journal of the ACM,* Vol. 46, Issue 5, pp. 604-632, 1999
- S. Brin, L. Page "The anatomy of a large-scale hypertextual Web search engine", *Proceedings of the seventh international conference on World Wide Web 7*: pp. 107-117, 1998
- K. Bryan, T. Leise, "The $25,000,000,000 eigenvector. The linear algebra behind Google", *SIAM Review*, Vol. 48, Issue 3, pp. 569-81. 2006.
- S. Vigna, "A critical, historical and mathematical review of centrality scores", SSMS Summer School, Santorini, 2012