DLI Accelerated Data Science Teaching Kit

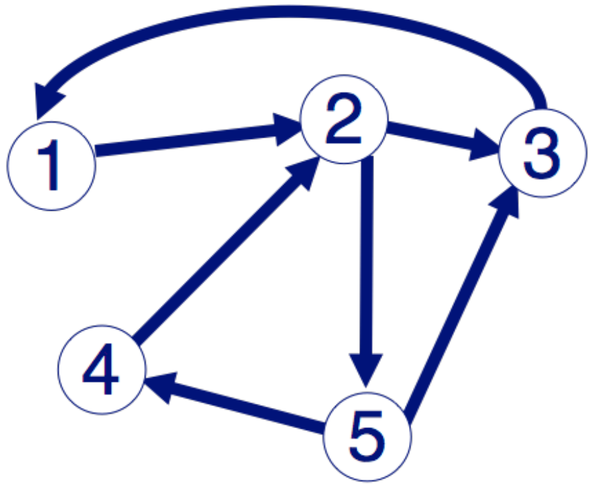
Lecture 17.4 - PageRank and Personalized PageRank

PageRank

Brin, Sergey and Lawrence Page (1998). Anatomy of a Large-Scale Hypertextual Web Search Engine. 7th Intl World Wide Web Conf.

PageRank: Problem

Given a directed graph, find its most interesting/central node

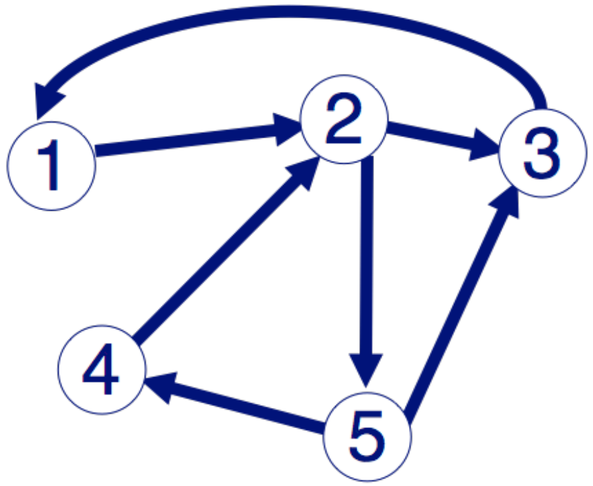


A node is important, if it is connected with important nodes (recursive, but OK!)

PageRank: Solution

Given a directed graph, find its most interesting/central node

Proposed solution: use random walk; most “popular” nodes are the ones with highest steady state probability (ssp)

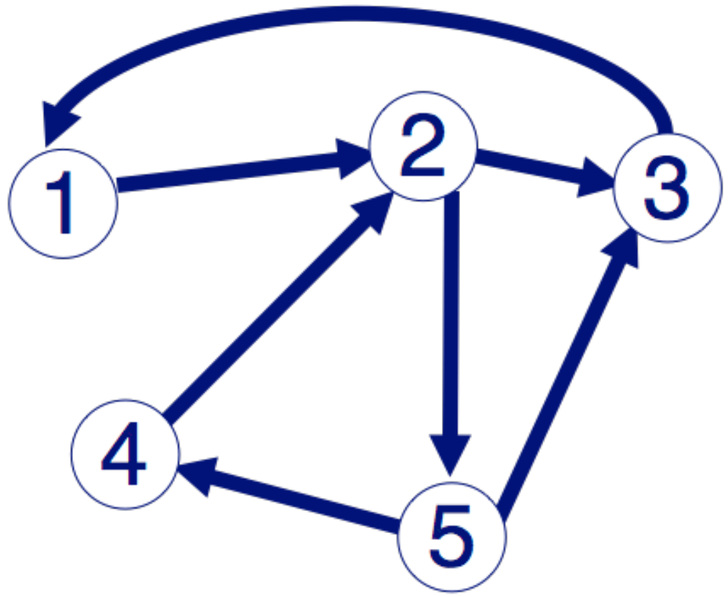


“state” = webpage

A node is important, if it is connected with important nodes (recursive, but OK!)

(Simplified) PageRank

Let B be the transition matrix: transposed, column-normalized



From / To

B

		1		
1			1	
	1/2			1/2
				1/2
	1/2			

p

p_1
p_2
p_3
p_4
p_5

=

=

p

p_1
p_2
p_3
p_4
p_5

How to compute SSP:

https://fenix.tecnico.ulisboa.pt/downloadFile/3779579688473/6.3.pdf

http://www.sosmath.com/matrix/markov/markov.html

(Simplified) PageRank

$$B\ p = 1 * p$$

Thus, p is the eigenvector that corresponds to the highest eigenvalue (=1, since the matrix is column-normalized)

Why does such a p exist?

- p exists if B is nxn, nonnegative, irreducible
  [Perron–Frobenius theorem]

(Simplified) PageRank

- In short: imagine a person randomly moving along the edges/links
- A node's PageRank score is the steady-state probability (ssp) of finding the person at that node

Full version of algorithm:

With occasional random jumps to any nodes

Why? To make the matrix irreducible.

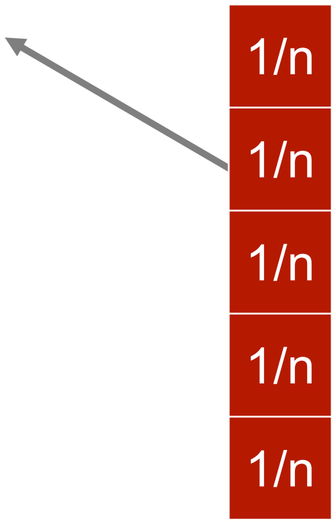
Irreducible = from any state (node), there's non-zero probability to reach any other state (node)

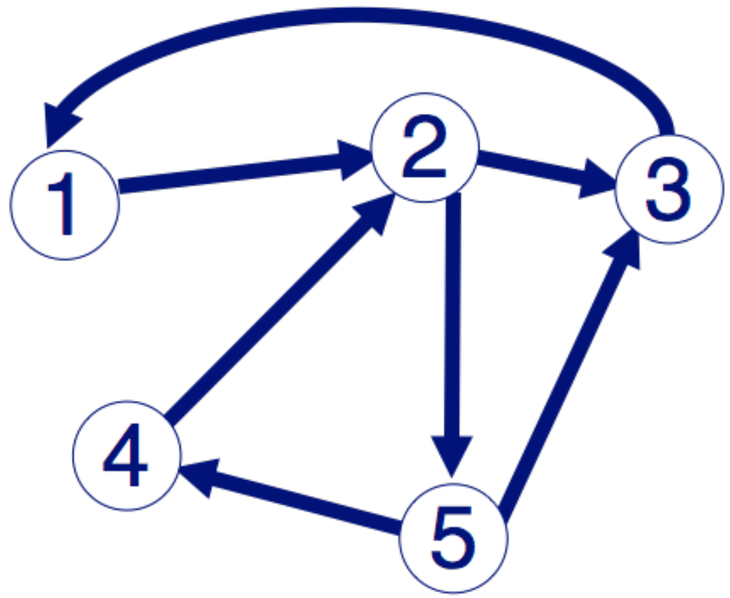
Full Algorithm

With probability 1-c, fly-out to a random node

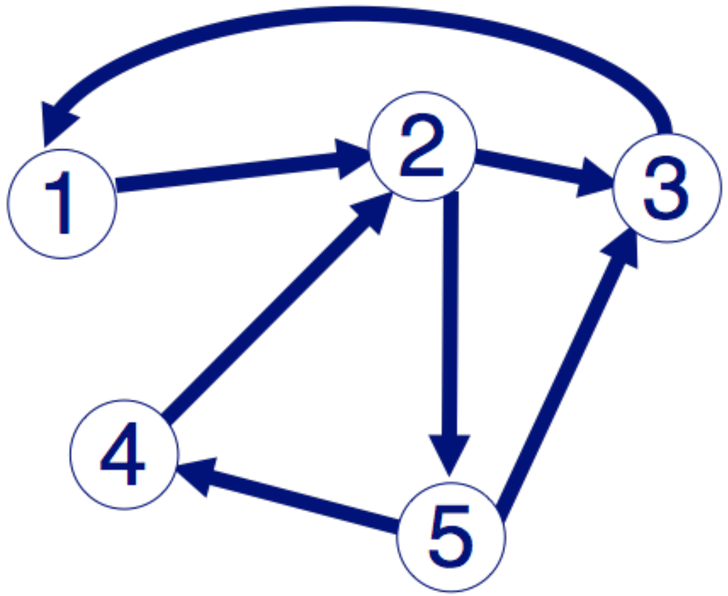
Then, we have

$$p = c\, B\, p + \frac{(1-c)}{n} 1$$





Run PageRank for Huge Matrix?



Use the power iteration method

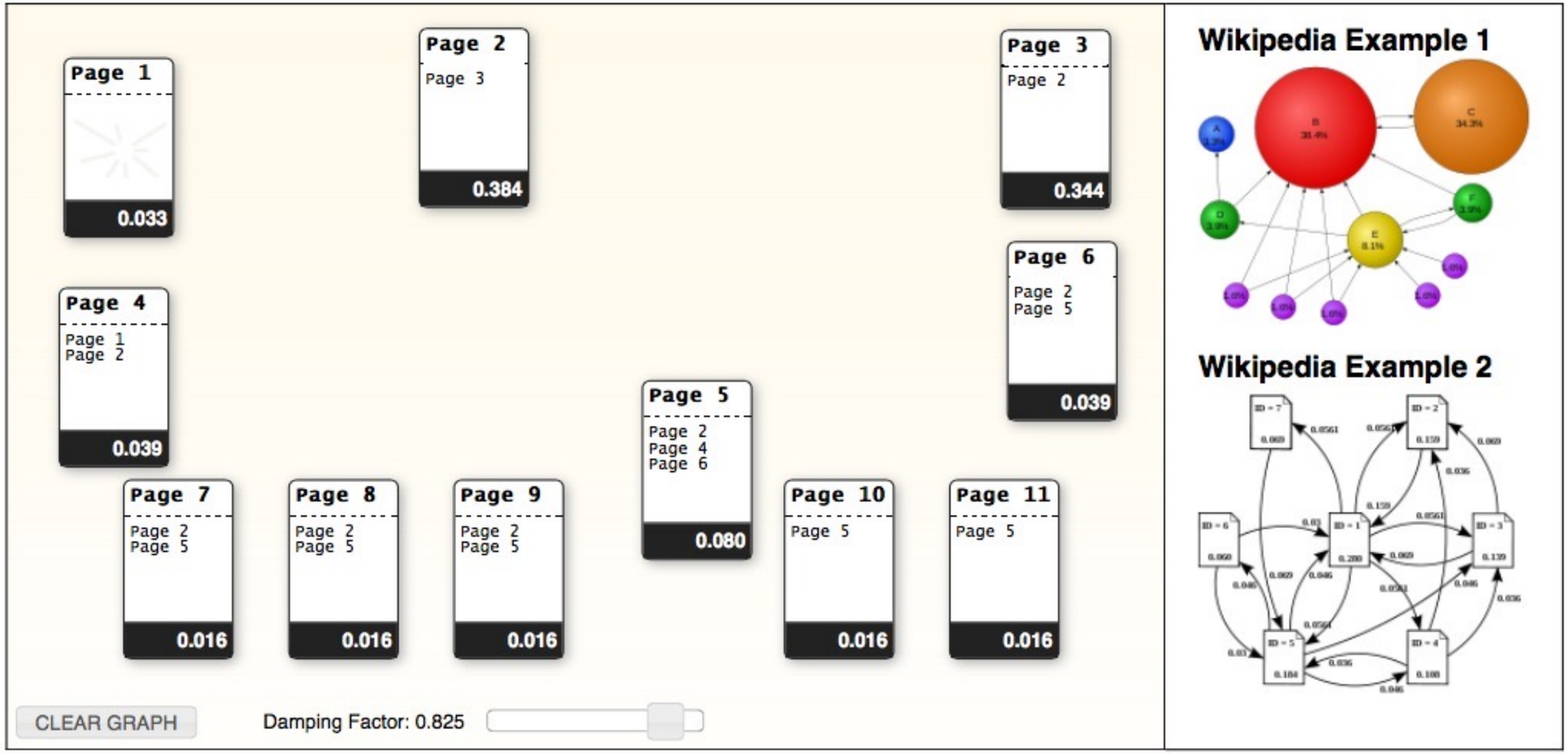
$$p' = c\,B\,p + \frac{(1-c)}{n}\,1$$

$$\begin{pmatrix} p'_1 \\ p'_2 \\ p'_3 \\ p'_4 \\ p'_5 \end{pmatrix} = c \begin{pmatrix} & & 1 & & \\ 1 & & & 1 & \\ & 1/2 & & & 1/2 \\ & & & & 1/2 \\ & 1/2 & & & \end{pmatrix} \begin{pmatrix} p_1 \\ p_2 \\ p_3 \\ p_4 \\ p_5 \end{pmatrix} + \frac{(1-c)}{n} \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$$

Can initialize this vector to any non-zero vector, e.g. all "1"s

http://en.wikipedia.org/wiki/Power_iteration

PageRank Explained with Javascript



Also great for checking the correctness of your PageRank Implementation.

http://www.cs.duke.edu/csed/principles/pagerank/

PageRank for Graphs (Generally)

You can run PageRank on any graphs

- All you need are the graph edges!

Should be in your algorithm “toolbox”

- Better than degree centrality
- Fast to compute for large graphs, runtime linear in the number of edges, O(E)

But can be “misled” (Google Bomb)

- How?

Personalized PageRank

Intuition: not all pages are equally relevant

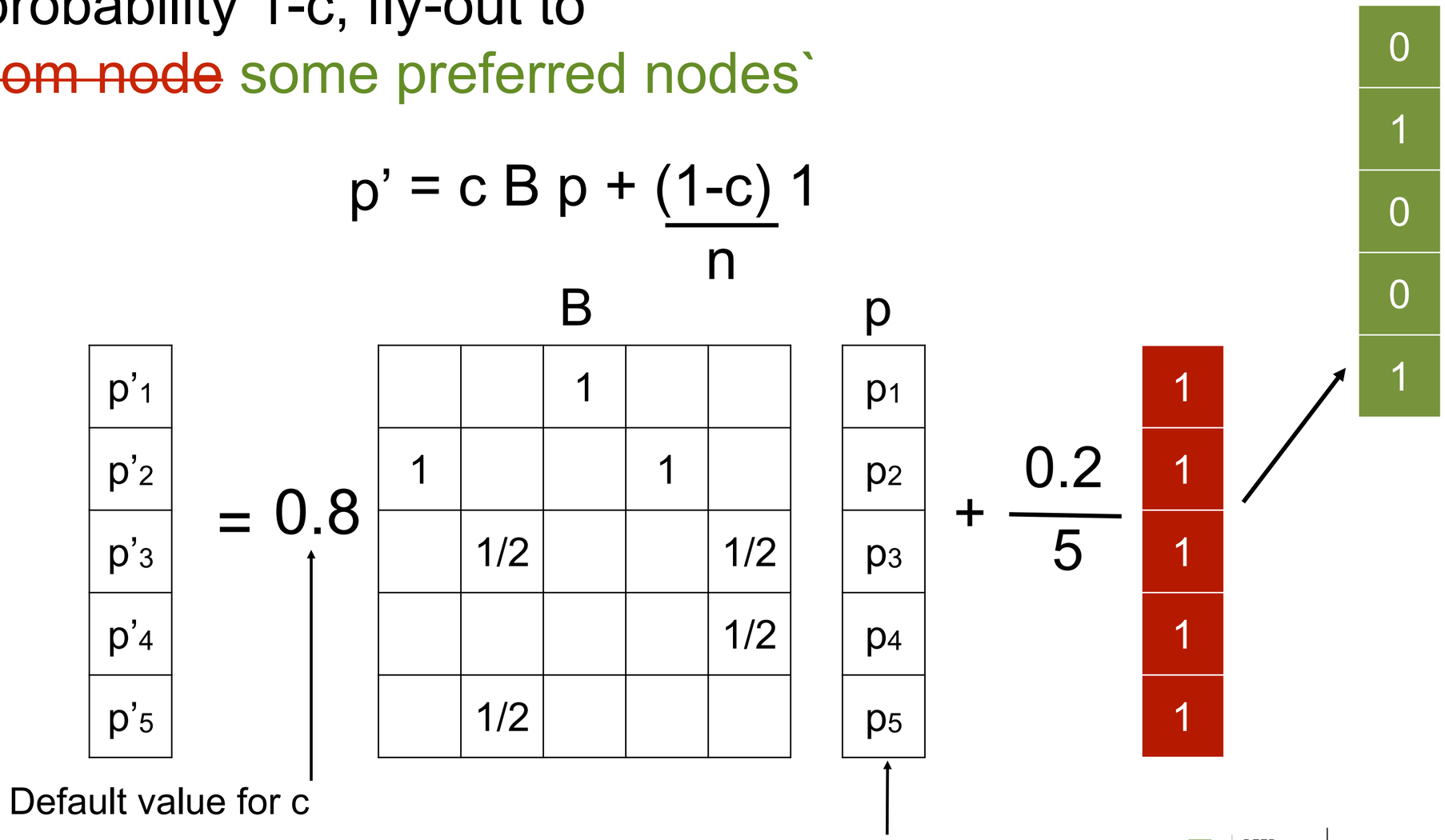
Goal: rank pages such that those more relevant to you will be ranked higher

How? Make just ONE small change to PageRank

Personalized PageRank

With probability 1-c, fly-out to
~~a random node~~ some preferred nodes`

$$p' = c\,B\,p + \frac{(1-c)}{n}\,1$$



Why Personalized PageRank?

For recommendation

- If I like webpage A, what else do I like?
- If I bought product A, what other products would I also buy?

Visualizing and interacting with large graphs

- Instead of visualizing every single nodes, visualize the most important ones

Very flexible — works on any graph

Related “Guilt-by-Association”/Diffusion Techniques

- Personalized PageRank
  (= Random Walk with Restart)

- “Spreading activation” or “degree of interest” in Human-Computer Interaction (HCI)

- Belief Propagation
  (powerful inference algorithm, for fraud detection, image segmentation, error-correcting codes, etc.)

Why are These Algorithms Popular?

- **Intuitive to interpret**
  Uses “network effect”, homophily
- **Easy to implement**
  Math is relatively simple (mainly matrix-vector multiplication)
- **Fast**
  Run time linear to #edges, or better
- **Probabilistic meaning**

DEEP
LEARNING
INSTITUTE

PRAIRIE VIEW
A&M UNIVERSITY

DLI Accelerated Data Science Teaching Kit

Thank You