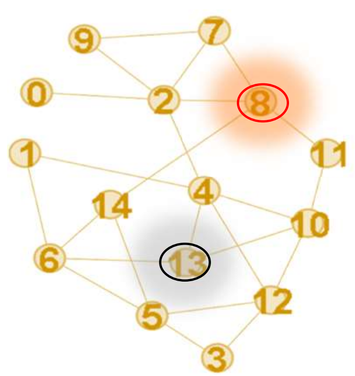
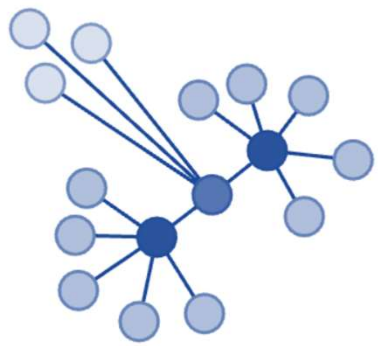
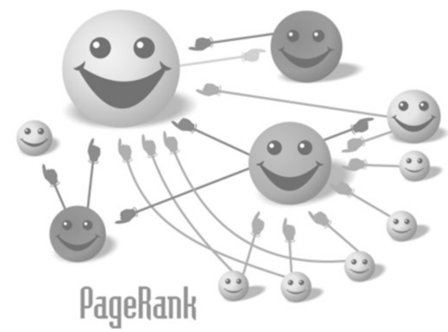
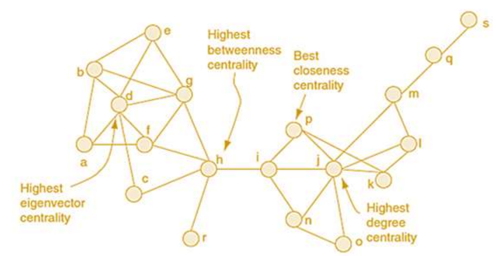
Prof. Ralucca Gera, rgera@nps.edu

Applied Mathematics Department,
Naval Postgraduate School









MA4404 Complex Networks

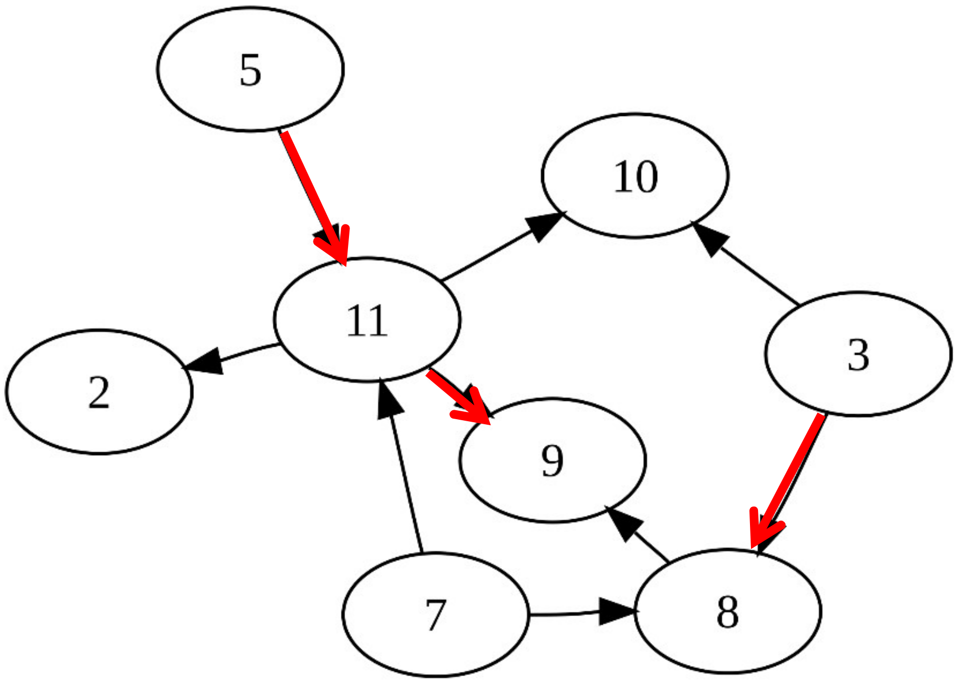
PageRank

Learning Outcomes

- Understand how PageRank is an extension of Katz and Eigenvector Centrality to directed graphs.
- Compute PageRank per node.
- Interpret the meaning of the values of PageRank.

Why PageRank?!

- Who knows how PageRank works? Guesses?
- In directed graphs: some in-degrees are zero.
- Fix: Katz centrality used a "free" weight of β
- New problem: should the weight of the following edges be the same:
  (11, 9),
  (5, 11),
  (3, 8)?
- How should we decide on the weight? Think about it while we're going through the slides.



Source: http://en.wikipedia.org/wiki/Directed_acyclic_graph

Introduction –web search

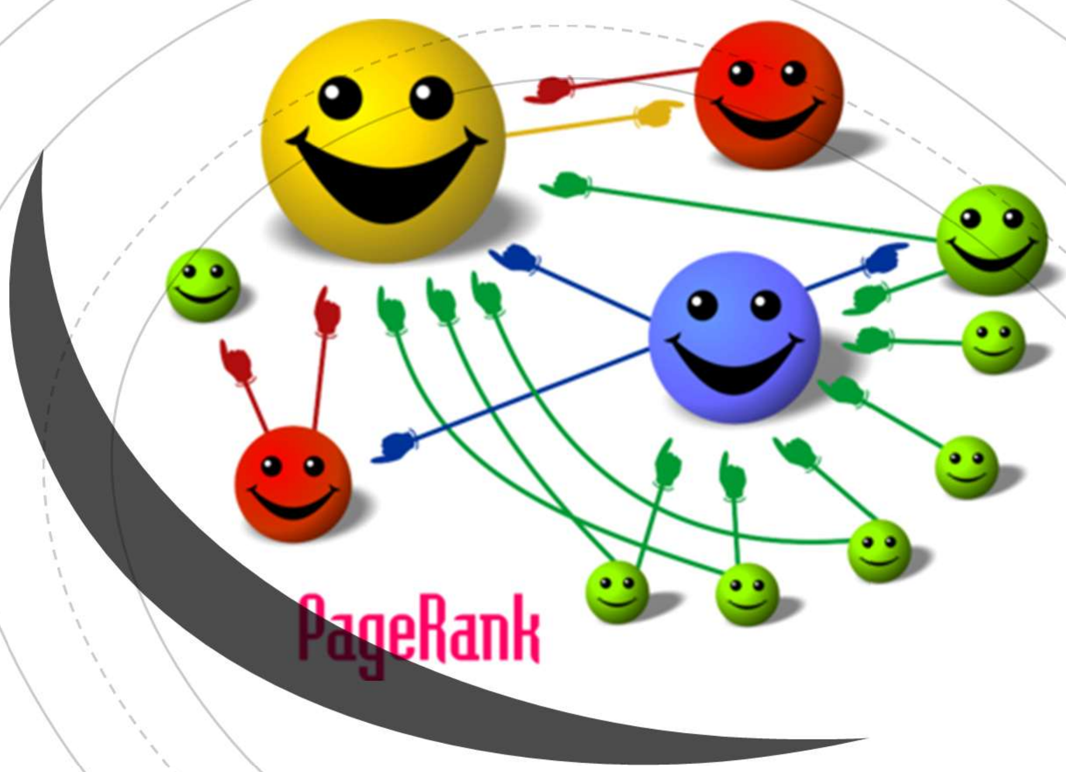
- Early search engines mainly compared content similarity of the query and the indexed pages. i.e.,
  - They use information retrieval methods, cosine similarity, TF-IDF, ...
- In the mid 1990's, it became clear that content similarity alone was no longer sufficient.
  - The number of pages grew rapidly in the mid 1990's.
    - How to choose only 30-40 pages and rank them suitably to present to the user?
  - Content similarity is easily spammed.
    - Webpage can repeat words and add related words to boost the rankings of his pages and/or to make the pages relevant to a large number of queries.

Introduction (cont ...)

- Starting around 1996, researchers began to work on the problem. They resorted to hyperlinks.
  - In 1997, Yanhong Li, Scotch Plains, NJ, created a hyperlink based search patent. The method uses **words** in anchor text of hyperlinks.
- Web pages on the other hand are connected through hyperlinks, which carry important information.
  - Some hyperlinks: organize information at the same site (anchors).
  - Other hyperlinks: point to pages from other Web sites. Such out-going hyperlinks often indicate an implicit conveyance of authority to the pages being pointed to.
- Those pages that are pointed to by many other pages are likely to contain authoritative information.

Introduction (cont …)

- During 1997-1998, two most influential hyperlink based search algorithms PageRank and HITS were published.
- Both algorithms exploit the hyperlinks of the Web to rank pages according to their levels of "prestige" or "authority".
  - **HITS (Section 7.5)**: Prof. Jon Kleinberg (Cornell University), at *Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, January 1998. ***(HITS stands for Hyperlink-Induced Topic Search)***
  - **PageRank (Section 7.4)**: Sergey Brin and Larry Page, PhD students from Stanford University, at *Seventh International World Wide Web Conference* (WWW7) in April, 1998.
- Which one have you heard of? Why?
- **HITS** is part of the **Ask** search engine (www.Ask.com).
- **PageRank** has emerged as the dominant link analysis model
  - due to its query-independence,
  - its ability to combat spamming, and
  - **Google's** huge business success.



Intuition behind PageRank

The PageRank Algorithm for WWW

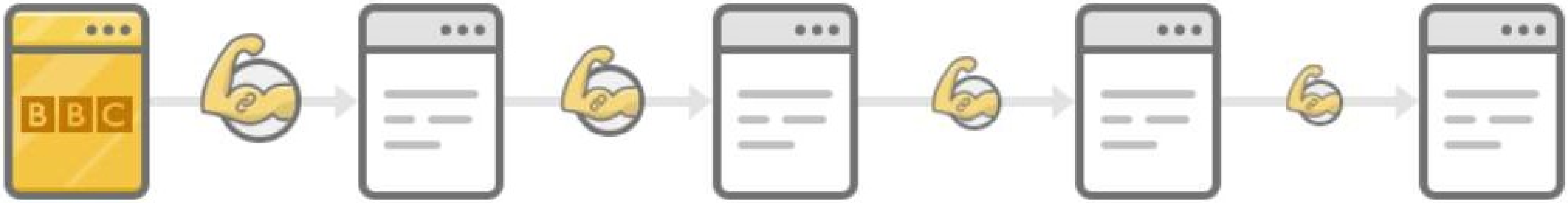
Sergey Brin and Larry Page
in 1998

(quitting their PhD programs at Stanford
to start Google)

- Invented the PageRank Algorithm to rank the returned key word searches
- PageRank is based on: A webpage is important if it is pointed to by other important pages.
- The algorithm was patented in 2001, and refined since.

PageRank: the intuitive idea

- PageRank relies on the democratic nature of the Web by using its vast link structure as an indicator of an individual page's value or quality.
- PageRank interprets a hyperlink from page *i* to page *j* as a vote, by page *i*, for page *j*.
- However, PageRank looks at more than the sheer number of votes; it also analyzes the page that casts the vote.
  - A vote casted by an “important” page *i* weighs more heavily and helps to make page *j* more "important." (like eigenvector and Katz)
  - Also, the vote of page *i* is shared among the pages that it points to, so page *j* gets a fraction of the vote.
    - How do we find that fraction? Think about it while we’re going through the slides



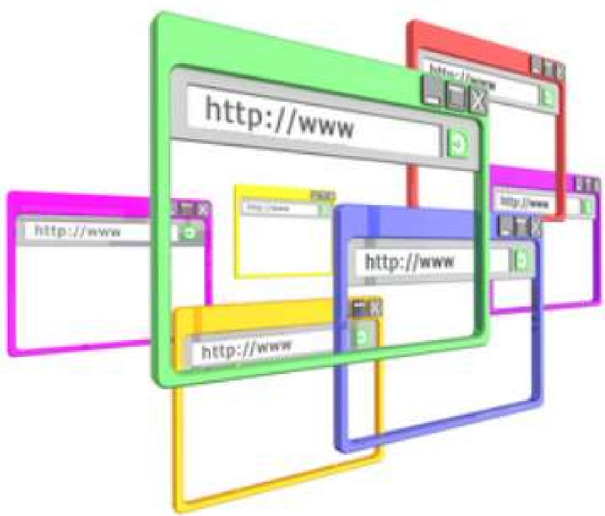
More specifically

- A hyperlink from a page to another page is an implicit transmission of authority to the target page.
  - ✓ The more in-links that a page *i* receives, the more prestige the page *i* has.
- Pages that point to page *i* also have their own prestige scores.
  - ✓ A page of a higher prestige pointing to *i* is more important than a page of a lower prestige pointing to *i*.
  - ✓ In other words, a page is important if it is pointed to by other important pages.

https://ahrefs.com/blog/google-pagerank/

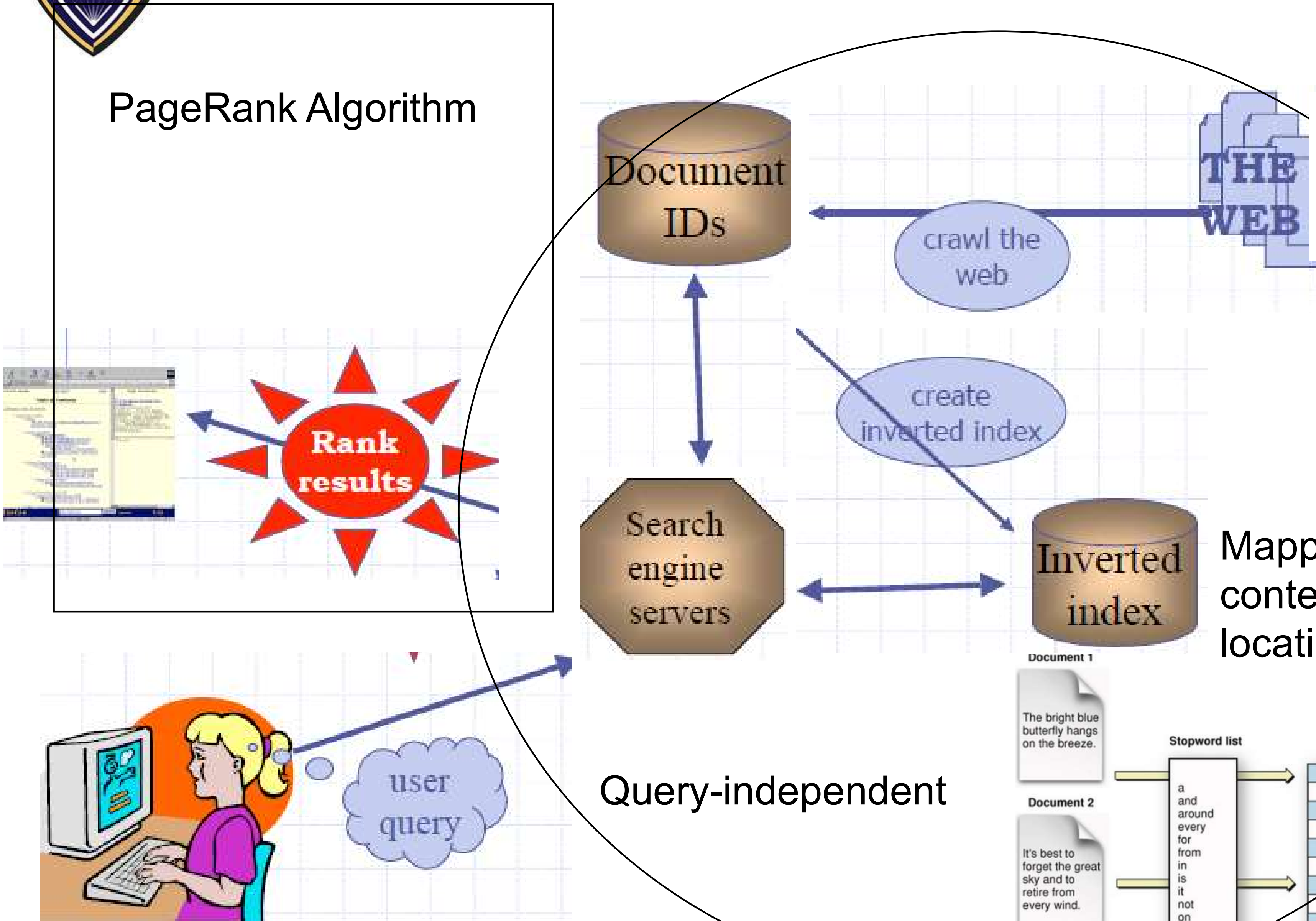
The web can be viewed as directed graph

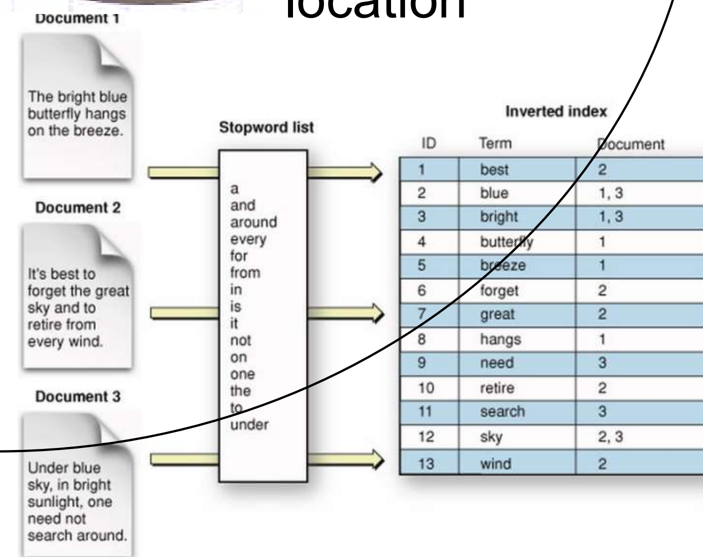
- The nodes or vertices are the web pages.
- The edges are the hyperlinks between websites
- This digraph has more than 10 billion vertices and it is growing every second!
- Google is useful because it ranks these outputs well, not because it finds all relevant pages



http://orleansmarketing.com/web-development1/microsites/#.VMX4xntHEqI

The web at a glance





Source: M. Ram Murty, Queen's University

Computing PageRank

$$x_i(t) = \sum_j A_{ij} x_j(t-1)$$

with the centrality at time t=0 being

$$x_j(0) = 1,\ \forall j$$

PageRank algorithm

- Eigenvector centrality: *i*'s Rank score, x_i, is the sum of the Rank scores x_j of all pages *j* that adjacent to *i* :

$$x_i \propto \sum_{(j,i)\in E} x_j$$

- Then Katz centrality adds **the teleportation** by adding a small weight edge to each node (using a weight of $\boldsymbol{\beta}$):

$$x_i = \frac{1}{\lambda_1} \sum_j A_{ij} x_j + \boldsymbol{\beta}$$

- BUT, since a page *j* may point to many other pages, its prestige score should be shared among these pages.

(For example, NP' main website pointing to many sites)

$$x_i = \alpha \sum_j A_{ij} \frac{x_j}{out\ deg_j} + \boldsymbol{\beta}$$

Matrix notation

- Let $\boldsymbol{x}$ be a *n*-dimensional column vector of PageRank values, i.e., $\boldsymbol{x} = (\boldsymbol{x_1}, \boldsymbol{x_2}, \dots, \boldsymbol{x_n})^T$.
- Let $\boldsymbol{A}$ be the adjacency matrix of our digraph with entries A_{ij}
- Then the PageRank centrality of node i is given by:

$$x_i = \alpha \sum_j A_{ij} \frac{x_j}{out \deg j} + \beta$$

or

$$\boldsymbol{x} = \alpha AD^{-1}\boldsymbol{x} + \beta$$

Where α is the damping factor, generally set for α = .85 (more on the next page).

Recall from eigenvector centrality:

$$\mathrm{A}\,\mathbf{x}(t) = \lambda_1 \mathbf{x}(t) \quad \text{or} \quad \mathbf{x}(t) = \frac{\mathbf{1}}{\boldsymbol{\lambda_1}} \mathrm{A}\,\mathbf{x}(t)$$

- Small α values (close to 0): the contribution given by paths longer than one hop is small, so centrality scores are mainly influenced by β (teleportation).
- Large α values (close to $\frac{\mathbf{1}}{\lambda_1}$): allows long paths to be devalued smoothly, and centrality scores influenced by the topology of G and less by the teleportation captured by β.
- Recommendation: choose $\alpha \in (0, \frac{\mathbf{1}}{\lambda_1})$, where the centrality diverges at α = $\frac{\mathbf{1}}{\lambda_1}$. The default is usually .85

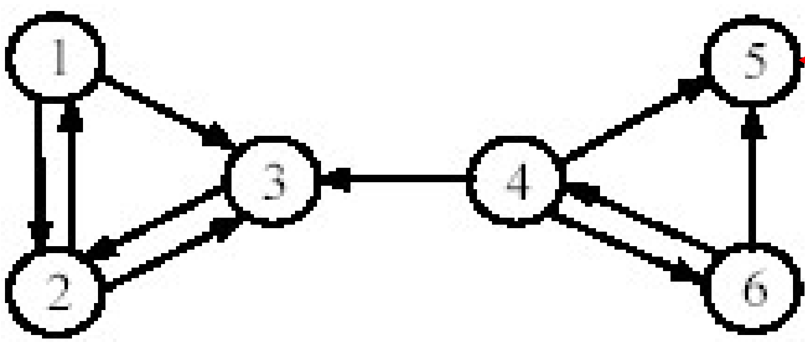
Updated Overview

Quality: what makes a node important (central)	**Mathematical Description**	**Appropriate Usage**	**Identification**
Lots of one-hop connections to high centrality vertices	A weighted degree centrality based on the weight of the neighbors	For example when the people you are connected to matter.	Eigenvector centrality $$C_i = \alpha \sum_j A_{ij} C_j$$ Where A is the in degree matrix
Lots of one-hop connections to high out-degree vertices	A weighted degree centrality based on the out degree of the neighbors	Directed graphs that are not strongly connected	Katz $C_i = \alpha \sum_j A_{ij} C_j + \beta$ Where β is some small weight for each node
As above but distribute the weight that a node has to the nodes it points to	$$\frac{C_j}{out\ \deg j}$$	As above but distributing the wealth of a node to the ones it points to	Page Rank: $$C_i = \alpha \sum_j A_{ij} \frac{C_j}{out\ \deg j} + \boldsymbol{\beta}$$ or $$\boldsymbol{x} = \alpha A D^{-1} \boldsymbol{x} + \beta \mathbf{1}$$

$$AD^{-1} = \begin{pmatrix} 0 & 1/2 & 0 & 0 & 1/6 & 0 \\ 1/2 & 0 & 1 & 0 & 1/6 & 0 \\ 1/2 & 1/2 & 0 & 1/3 & 1/6 & 0 \\ 0 & 0 & 0 & 0 & 1/6 & 1/2 \\ 0 & 0 & 0 & 1/3 & 1/6 & 1/2 \\ 0 & 0 & 0 & 1/3 & 1/6 & 0 \end{pmatrix}$$

An example using the Adjacency and Diagonal Matrices

An example as just described:



Problem vertex
(no outgoing links)

in-degree matrix

$$A = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

each row shows the in degree

each column shows the out degree

$$x_i = \alpha \sum_j A_{ij} \frac{x_j}{out \deg j} + \beta$$

or

$$\boldsymbol{x} = \alpha \, A D^{-1} \boldsymbol{x} + \beta \mathbf{1}$$

Recall that the problem with vertices with indegree = 0 was solved by using β.

Is the formula above well defined?

If not, how could we fix the formula or the matrix?

How can we fix the problem?

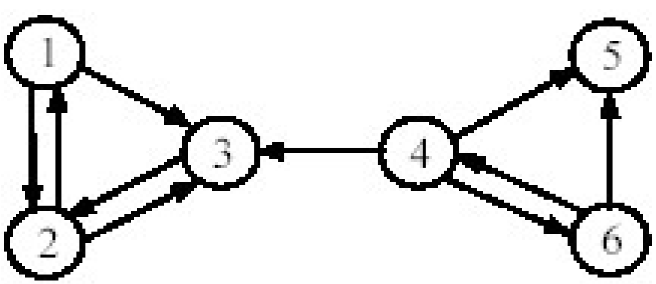
1. Remove those pages with no out-links during the PageRank computation as these pages do not affect the ranking of any other page directly (these pages will get outgoing links in the future).
2. Add a complete set of outgoing links from each such page *i* to all the pages on the Web.

each column shows the out degree

in-degree matrix

each row shows the in degree

The second choice is used in PR since matrix may get updated



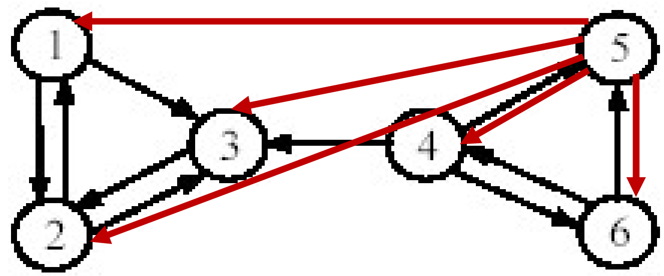
$$A = \begin{pmatrix} 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 \end{pmatrix}$$

How can we fix the out degree = 0?

$$\boldsymbol{A} = \begin{pmatrix} 0 & 1 & 0 & 0 & \boxed{1} & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 \end{pmatrix}$$

in-degree matrix

$$\boldsymbol{x} = \alpha A D^{-1} \boldsymbol{x} + \beta \mathbf{1}$$



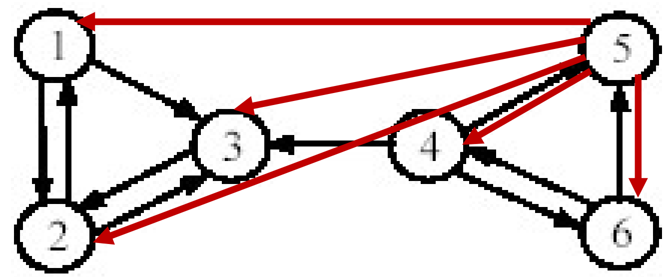
$$\boldsymbol{D}^{-1} = \begin{pmatrix} 1/2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1/2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1/1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/3 & 0 & 0 \\ 0 & 0 & 0 & 0 & \boxed{1/6} & 0 \\ 0 & 0 & 0 & 0 & 0 & 1/2 \end{pmatrix}$$

Inverse of the out-degree matrix

PR centrality formula is well defined

By multiplying them we obtain the matrix that captures: $\boldsymbol{x} = \alpha AD^{-1}\boldsymbol{x} + \beta\mathbf{1}$

1. The in and out degree per vertex
2. Divides the centrality of each vertex by its degree



The contribution of node 5 is insignificant, and the formula is now well defined

out-degree matrix

in-degree matrix

$$AD^{-1} = \begin{pmatrix} 0 & 1/2 & 0 & 0 & \boxed{1/6} & 0 \\ 1/2 & 0 & 1 & 0 & 1/6 & 0 \\ 1/2 & 1/2 & 0 & 1/3 & 1/6 & 0 \\ 0 & 0 & 0 & 0 & 1/6 & 1/2 \\ 0 & 0 & 0 & 1/3 & 1/6 & 1/2 \\ 0 & 0 & 0 & 1/3 & 1/6 & 0 \end{pmatrix}$$

Transition probability matrix

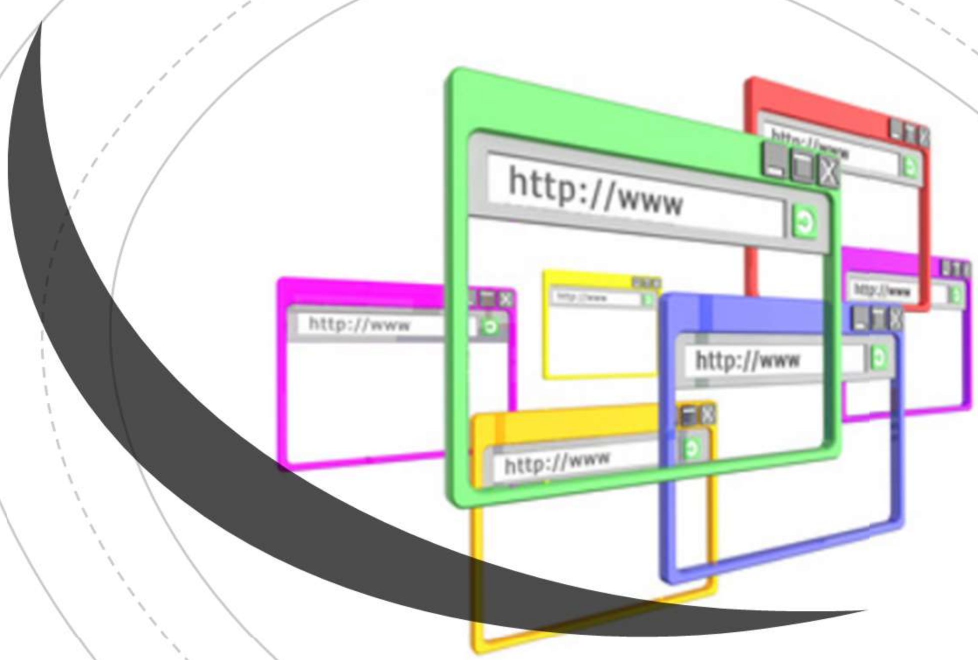
- This modified AD^{-1} matrix is called the state transition probability matrix. Denote its entries by p_{ij} :

$$\boldsymbol{AD^{-1}} = \begin{pmatrix} p_{11} & p_{12} & \cdot & \cdot & \cdot & p_{1n} \\ p_{21} & p_{22} & \cdot & \cdot & \cdot & p_{2n} \\ \cdot & \cdot & & & & \cdot \\ \cdot & \cdot & & & & \cdot \\ \cdot & \cdot & & & & \cdot \\ p_{n1} & p_{n2} & \cdot & \cdot & \cdot & p_{nn} \end{pmatrix}$$

- p_{ij} represents the transition probability that the surfer in state *i* (page *i*) will move to state *j* (page *j*).
- An extra example in the backup slides of this PPT deck.

Updated Overview

Quality: what makes a node important (central)	Mathematical Description	Appropriate Usage	Identification
Lots of one-hop connections to high centrality vertices	A weighted degree centrality based on the weight of the neighbors	For example when the people you are connected to matter.	Eigenvector centrality $$C_i = \alpha \sum_j A_{ij} C_j$$ Where A is the in degree matrix
Lots of one-hop connections to high out-degree vertices	A weighted degree centrality based on the out degree of the neighbors	Directed graphs that are not strongly connected	Katz $C_i = \alpha \sum_j A_{ij} C_j + \beta$ Where β is some small weight for each node
As above but distribute the weight that a node has to the nodes it points to	$$\frac{C_j}{out \deg j}$$	As above but distributing the wealth of a node to the ones it points to	Page Rank: $$C_i = \alpha \sum_j A_{ij} \frac{C_j}{out \deg j} + \boldsymbol{\beta}$$ Where outdeg j = max{1, out degree of node j}, or $$\boldsymbol{x} = \alpha A D^{-1} \boldsymbol{x} + \beta \mathbf{1}$$



Final Comments

Some comments

- Newman's book gives: $x_i = \boldsymbol{\alpha} \sum_j A_{ij} \frac{x_j}{out\ \mathrm{de}} + \boldsymbol{\beta}$

where α is called the **damping factor** which can be set to between 0 and 1(or the largest eigenvalue of *A*).

- And the formula in the original PageRank is:

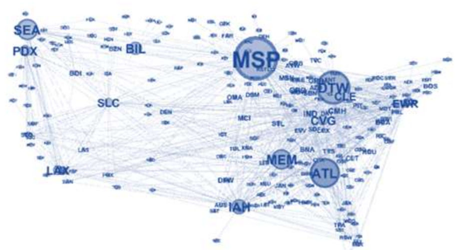
$$x_i = \boldsymbol{d} \sum_j A_{ij} \frac{x_j}{out\ \deg j} + (1 - d)$$

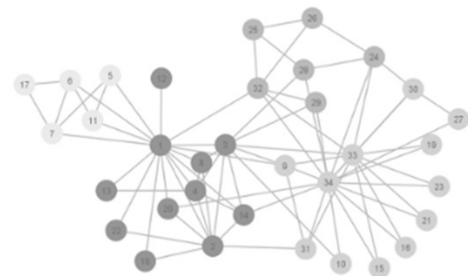
where ***d*** is the **damping factor (*d* = 0.85 as default**)

- Gephi: the default value for $\boldsymbol{\beta}$ is the **probability = 0.85** and Epsilon is the criteria for eigenvector convergence based on the power method

Final Points on PageRank

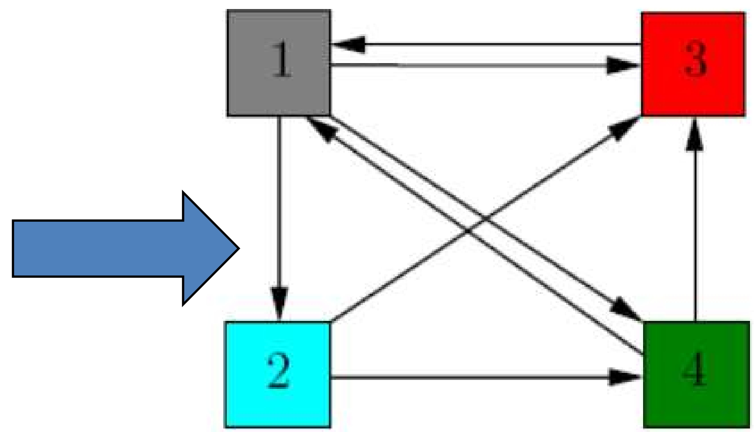
- **Fighting spam**.
  - A page is important if the pages pointing to it are important.
  - Since it is not easy for Web page owner to add in-links into his/her page from other important pages, it is thus not easy to influence PageRank.
- **PageRank is a global measure and is query independent**.
  - The values of the PageRank algorithm of all the pages are computed and saved off-line rather than at the query time => fast
- **Criticism**:
  - There are companies that can increase your PageRank by adding it to a cluster and increasing its indegree
  - It cannot not distinguish between pages that are authoritative in general and pages that are authoritative on the query topic.
    - But it works based on the keyword search





Back up slides: one smaller example

A 4-website Internet

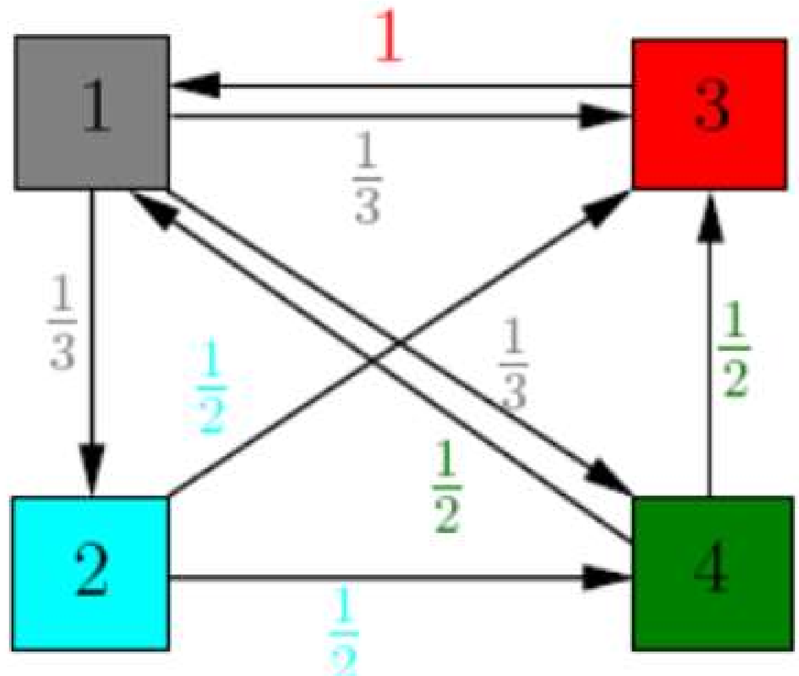


Source: http://www.math.cornell.edu/~mec/Winter2009/RalucaRemus/Lecture3/lecture3.html

A 4-website Internet

p_{ij} represents the transition probability that the surfer on page *j* will move to page *i*:

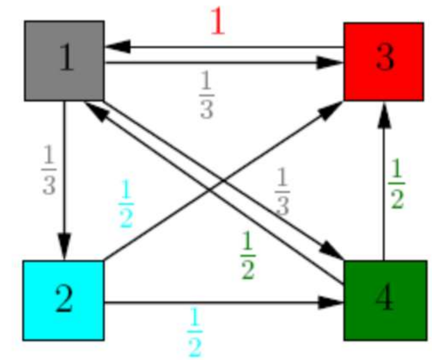
$$AD^{-1} = (p_{ij}) = \begin{pmatrix} 0 & 0 & 1 & 1/2 \\ 1/3 & 0 & 0 & 0 \\ 1/3 & 1/2 & 0 & 1/2 \\ 1/3 & 1/2 & 0 & 0 \end{pmatrix}$$



Source: http://www.math.cornell.edu/~mec/Winter2009/RalucaRemus/Lecture3/lecture3.html

A 4-website Internet

Random surfer: each page has equal probability ¼ to be chosen as a starting point.



$$v = \begin{pmatrix} 0.25 \\ 0.25 \\ 0.25 \\ 0.25 \end{pmatrix}, \quad Av = \begin{pmatrix} 0.37 \\ 0.08 \\ 0.33 \\ 0.20 \end{pmatrix}, \quad A^2 v = A(Av) = A \begin{pmatrix} 0.37 \\ 0.08 \\ 0.33 \\ 0.20 \end{pmatrix} = \begin{pmatrix} 0.43 \\ 0.12 \\ 0.27 \\ 0.16 \end{pmatrix}$$

$$A^3 v = \begin{pmatrix} 0.35 \\ 0.14 \\ 0.29 \\ 0.20 \end{pmatrix}, \quad A^4 v = \begin{pmatrix} 0.39 \\ 0.11 \\ 0.29 \\ 0.19 \end{pmatrix}, \quad A^5 v = \begin{pmatrix} 0.39 \\ 0.13 \\ 0.28 \\ 0.19 \end{pmatrix}$$

$$A^6 v = \begin{pmatrix} 0.38 \\ 0.13 \\ 0.29 \\ 0.19 \end{pmatrix}, \quad A^7 v = \begin{pmatrix} 0.38 \\ 0.12 \\ 0.29 \\ 0.19 \end{pmatrix}, \quad A^8 v = \begin{pmatrix} 0.38 \\ 0.12 \\ 0.29 \\ 0.19 \end{pmatrix}$$

$$AD^{-1} = (p_{ij}) = \begin{pmatrix} 0 & 0 & 1 & 1/2 \\ 1/3 & 0 & 0 & 0 \\ 1/3 & 1/2 & 0 & 1/2 \\ 1/3 & 1/2 & 0 & 0 \end{pmatrix}$$

The probability that page ***i*** will be visited after *k* steps (i.e. the random surfer ending up at page ***i***) is equal to i^{th} entry of $A^k x$.

the sequences of iterates $v, Av, \ldots, A^k v$ tends to the equilibrium value $v^* = \begin{pmatrix} 0.38 \\ 0.12 \\ 0.29 \\ 0.19 \end{pmatrix}$

call this the PageRank vector of our web graph.

Simplification for this example: No β was involved since id $i > 0$, for all i

Source: http://www.math.cornell.edu/~mec/Winter2009/RalucaRemus/Lecture3/lecture3.html