CAP 5768: Introduction to Data Science

Giri NARASIMHAN

www.cis.fiu.edu/~giri/teach/5768.html

PageRank & Link Analysis

Early Search Engines

- Crawl the web, collect terms, build inverted index (term to URL mapping)
- Spammers found tricks to beat the system
  - Add irrelevant terms to URL (Term Spam)
  - Copy top hit URL to your page

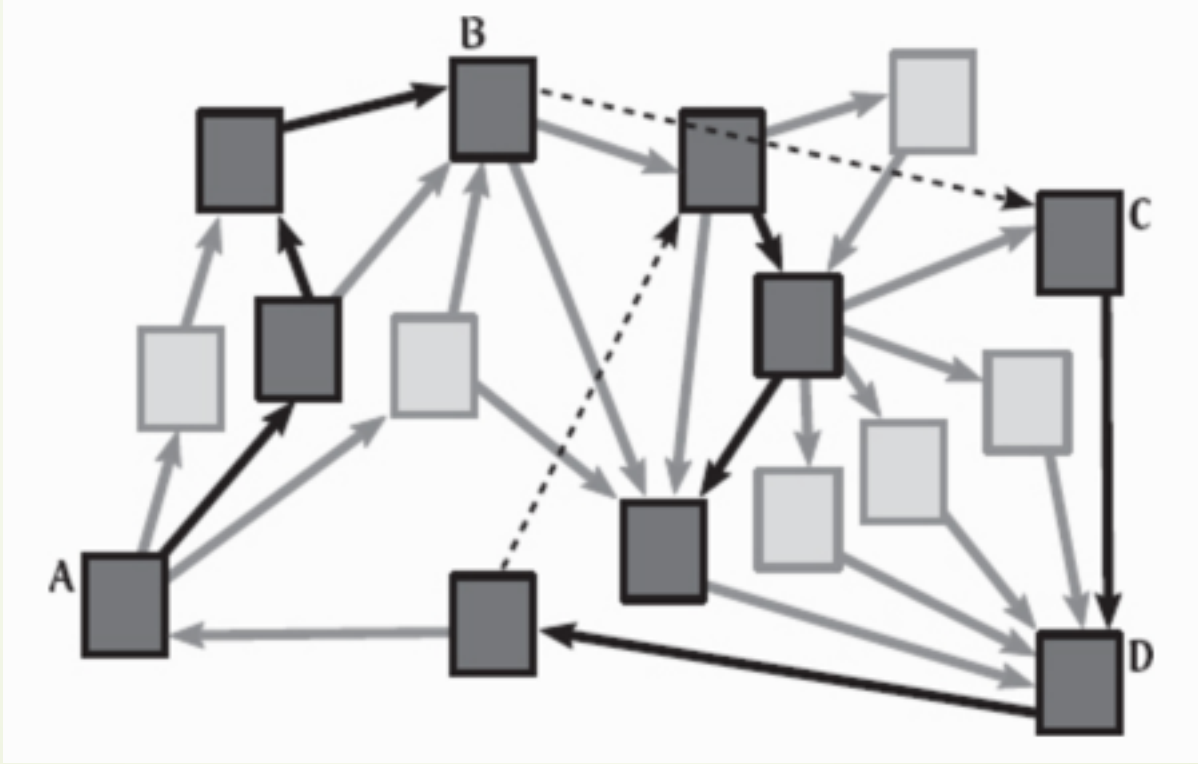
Importance of Web Pages

- If more links into X, then X is important
- If important pages link to X, then X is important
  - Chicken and Egg problem for importance
- Random Walk Idea
  - Simulate random walk & count # of recurring visits
  - Spam farm problem to trap random walker

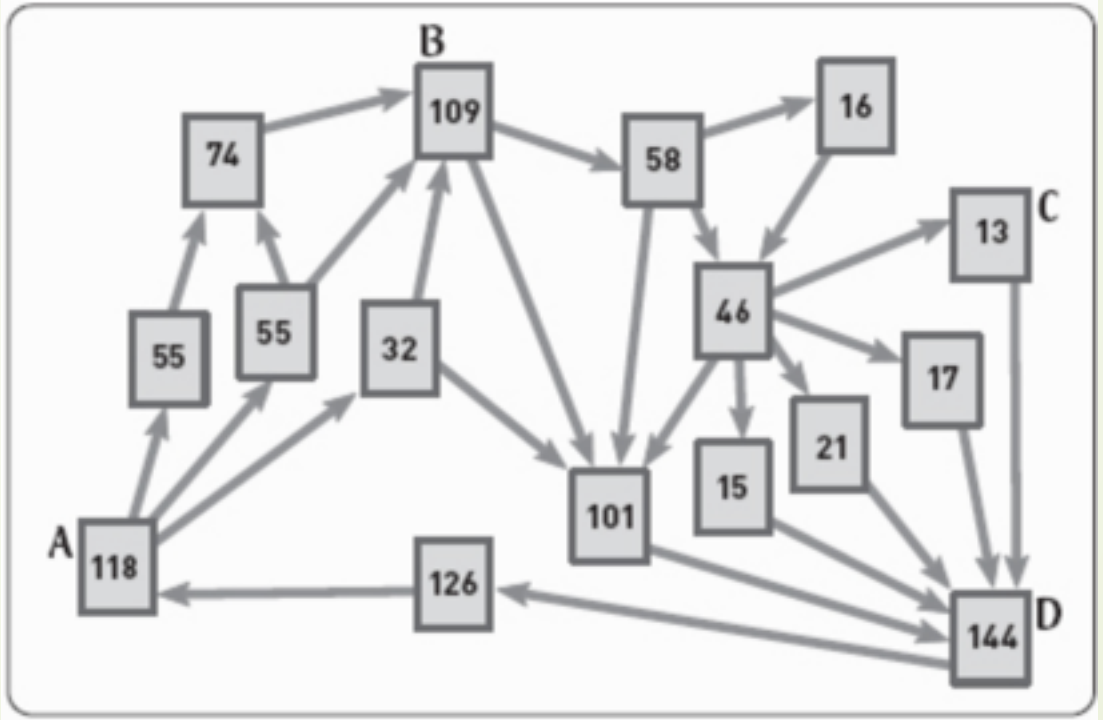
Random Walks & PageRank

- Perform random walks
- Importance of page is proportional to how often you visit a node
- During web search, prefer to report important pages

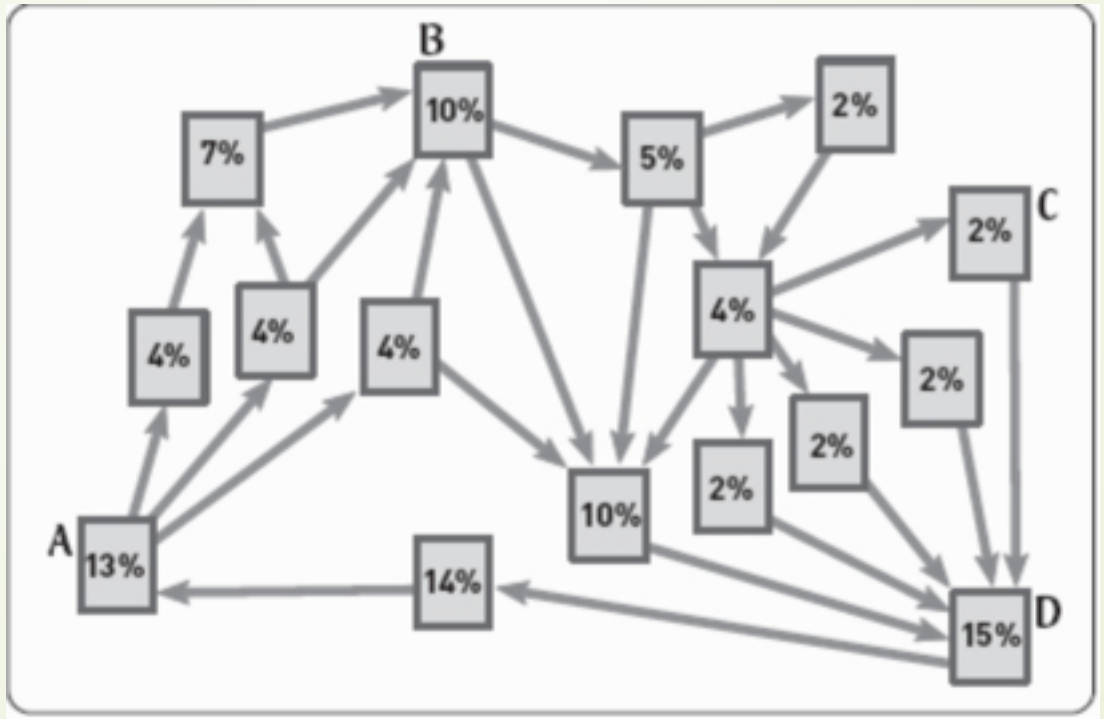
Example



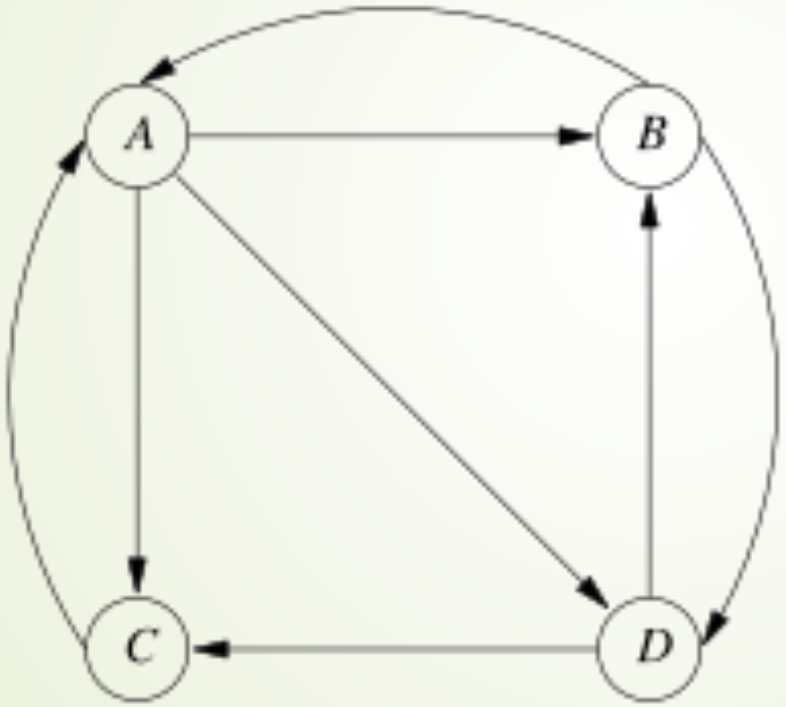
Example



Example



Random Walks



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

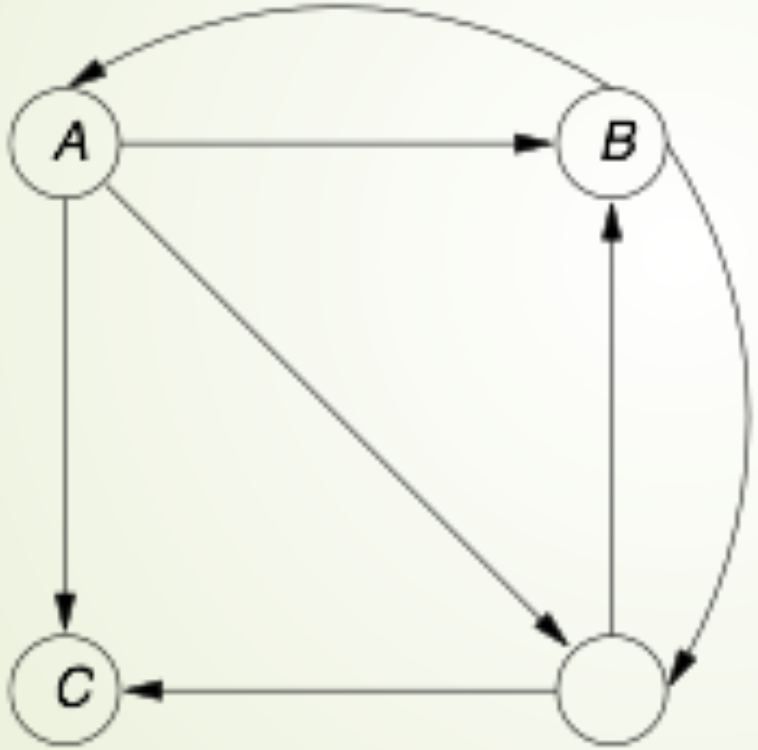
$$\begin{bmatrix} 3/9 \\ 2/9 \\ 2/9 \\ 2/9 \end{bmatrix}$$

Computing the Stationary Prob

$$M = \begin{bmatrix} 0 & 1/2 & 1 & 0 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 1/2 & 0 & 0 \end{bmatrix} \qquad \mathbf{v} = M\mathbf{v}$$

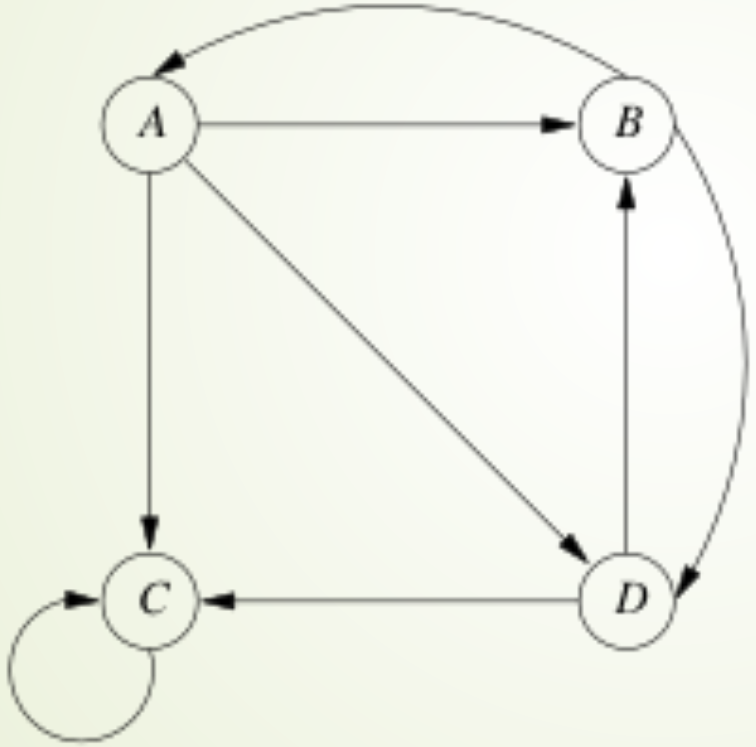
$$\begin{bmatrix} 1/4 \\ 1/4 \\ 1/4 \\ 1/4 \end{bmatrix} \begin{bmatrix} 9/24 \\ 5/24 \\ 5/24 \\ 5/24 \end{bmatrix} \begin{bmatrix} 15/48 \\ 11/48 \\ 11/48 \\ 11/48 \end{bmatrix} \begin{bmatrix} 11/32 \\ 7/32 \\ 7/32 \\ 7/32 \end{bmatrix} \cdots \begin{bmatrix} 3/9 \\ 2/9 \\ 2/9 \\ 2/9 \end{bmatrix}$$

Random Walks w/ Dead Ends



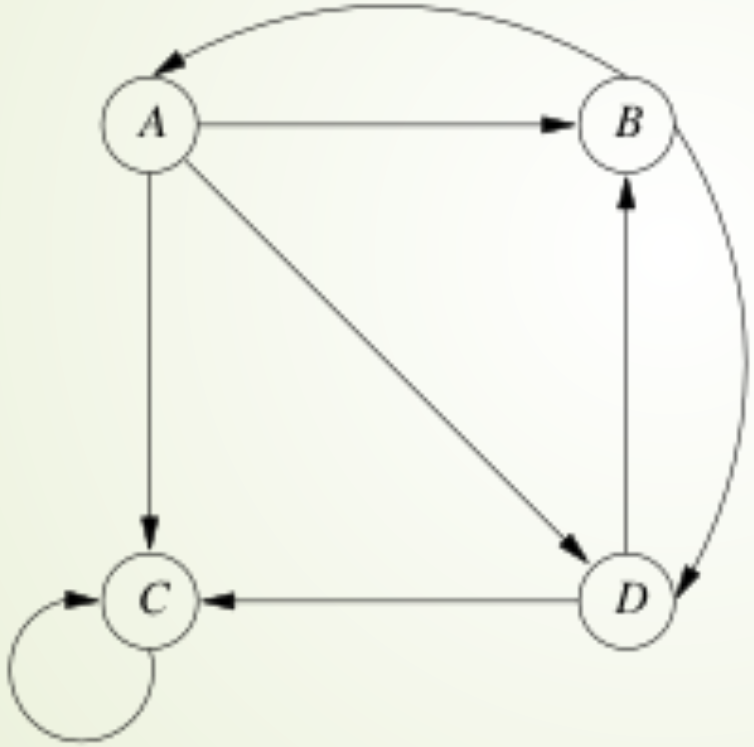
$$\cdot \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

Rand Walks w/o Dead Ends



$$\begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}$$

Rand Walk w/ Teleporting (0.8)

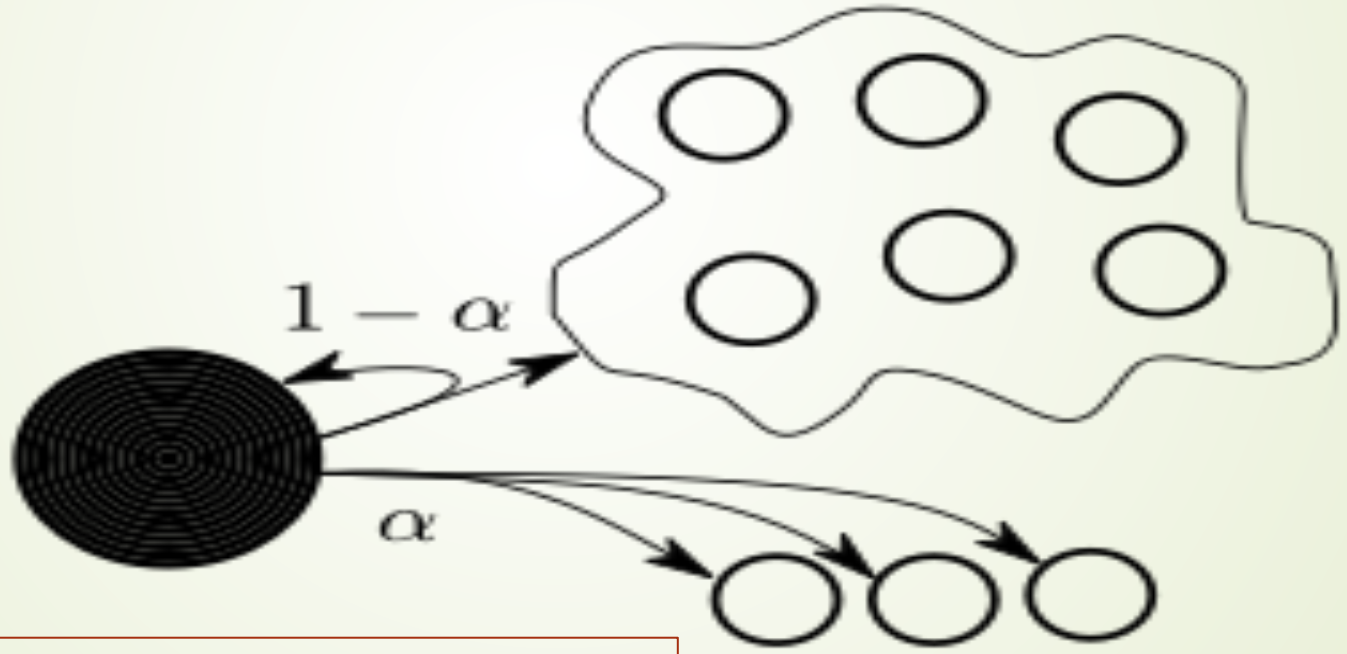


$$\begin{bmatrix} 15/148 \\ 19/148 \\ 95/148 \\ 19/148 \end{bmatrix}$$

PageRank

- Essential component of Google search engine
- PageRank allows efficient and stable prioritization of search results
- Vote of Confidence Principle
  - PageRank of a web page will be high if it is linked to other highly ranked pages
- Random Walk / Markov Chain Analogy
  - Which pages are most visited on a random walk?
- Teleportation Analogy
  - Models when a user jumps next to an unlinked page

Random Walk + Teleportation Analogy



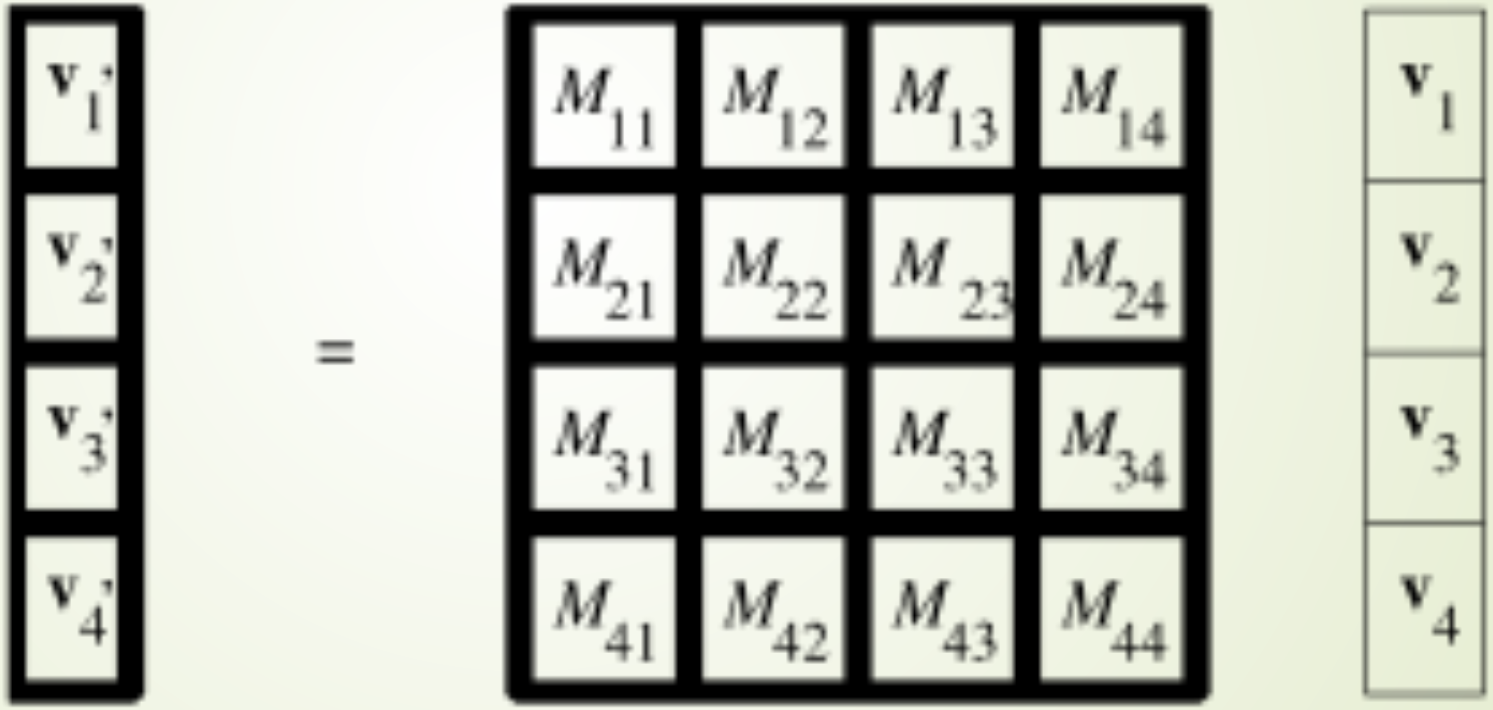
$$\mathbf{v}' = \beta M\mathbf{v} + (1-\beta)\mathbf{e}/n$$

Implementing PageRank

$$\mathbf{v}' = \beta M \mathbf{v} + (1 - \beta)\mathbf{e}/n$$

- Even v may not fit in memory
- Use MapReduce

Partitioning for MapReduce



Impact of teleportation parameter

10 Wikipedia pages with highest PageRank [Gleich, '09]

$\alpha = 0.50$	$\alpha = 0.85$	$\alpha = 0.99$
United States	United States	C:Contents
C:Living people	C:Main topic classif.	C:Main topic classif.
France	C:Contents	C:Fundamental
Germany	C:Living people	United States
England	C:Ctgs. by country	C:Wikipedia admin.
United Kingdom	United Kingdom	P:List of portals
Canada	C:Fundamental	P:Contents/Portals
Japan	C:Ctgs. by topic	C:Portals
Poland	C:Wikipedia admin.	C:Society
Australia	France	C:Ctgs. by topic

Applications of PageRank

- **Clustering** [Andersen '06] [reset to same page]
- **Sports Ranking** [Govan '08]
- **Bioinformatics** – GeneRank [Morrison '05], ProteinRank [Freschi '07]
- **Network Alignment** [Singh '07]
- **Literature** – BookRank; Bibliometrics – CiteRank, AuthorRank, TimedPageRank;
- **Information Systems** – PopRank, FactRank, ObjectRank, FolkRank
- **Recommender Systems** – ItemRank
- **Social Networks** – BuddyRank, TwitterRank
- **Web** – HostRank, DirRank, TrustRank, BadRank, VisualRank

Mathematics of PageRank

- Given: P_{ij} = prob of transition from **j** to **i**
- PageRank is given by solution **x** to equation
  - $(aP + (1-a)ve^T)x = x$
  - Thus **x** is eigenvector of a certain matrix
  - Alternative equation: $(I-aP)x = (1-a)v$
  - Eigenvectors can be computed using
  - ge e al. used an iterati powering method:
    - (1-a)v
    - 3656 ite -16 error

Random Walk

Teleportation

Variants of PageRank: Localized

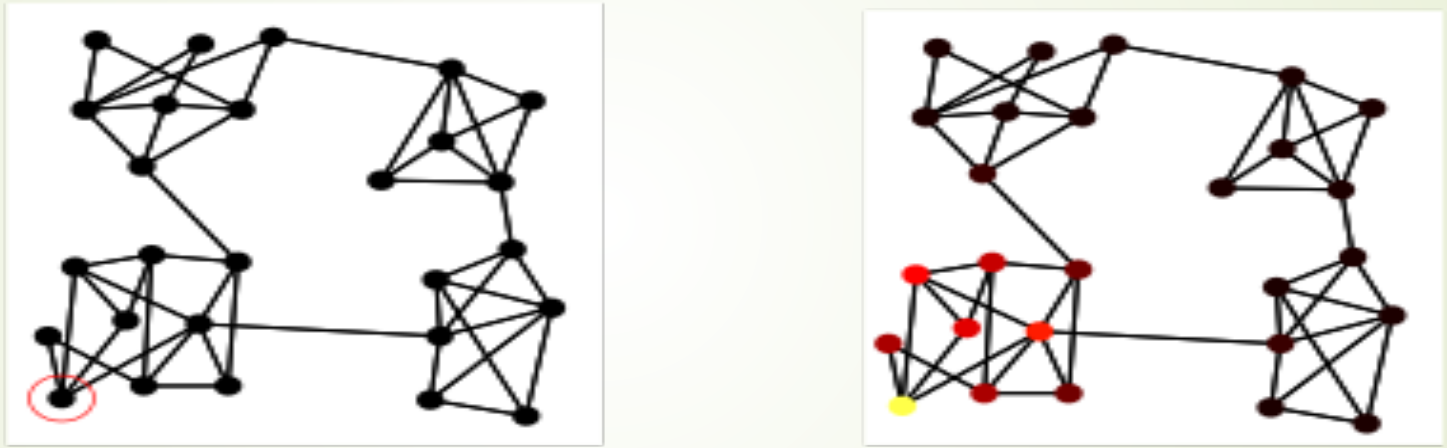


FIG. 2.1. *An illustration of the empirical properties of localized PageRank vectors with teleportation to a single node in an isolated region. In the graph at left, the teleportation vector is the single circled node. The PageRank vector is shown as the node color in the right figure. PageRank values remain high within this region and are nearly zero in the rest of the graph. Theory from Andersen et al. [2006] explains when this property occurs.*

Topic-Sensitive PageRank

- Teleport to a URL with same topic

Variants of PageRank

$$\mathbf{h} = \lambda\mu LL^{\mathrm{T}}\mathbf{h}.$$

$$\mathbf{a} = \lambda\mu L^{\mathrm{T}}L\mathbf{a}.$$

- Hubs & Authorities
  - Authorities are nodes with information
    - E.g., Course webpage
  - Hubs provide links to authorities
    - E.g., list of courses offered

$$\mathbf{a} = \mu L^{\mathrm{T}}\mathbf{h},$$

- Good Authorities linked from Good Hubs
- Good Hubs link to Good Authorities

$$\mathbf{h} = \lambda L\mathbf{a},$$

Computing h and a

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

$$\mathbf{h} \qquad L^{\mathrm{T}}\mathbf{h} \qquad \mathbf{a} \qquad L\mathbf{a} \qquad \mathbf{h}$$

$$\begin{bmatrix} 1/2 \\ 5/3 \\ 5/3 \\ 3/2 \\ 1/6 \end{bmatrix} \quad \begin{bmatrix} 3/10 \\ 1 \\ 1 \\ 9/10 \\ 1/10 \end{bmatrix} \quad \begin{bmatrix} 29/10 \\ 6/5 \\ 1/10 \\ 2 \\ 0 \end{bmatrix} \quad \begin{bmatrix} 1 \\ 12/29 \\ 1/29 \\ 20/29 \\ 0 \end{bmatrix}$$

$$L^{\mathrm{T}}\mathbf{h} \qquad \mathbf{a} \qquad L\mathbf{a} \qquad \mathbf{h}$$

Variants of PageRank

- Reverse PageRank (follow inlinks, not outlinks)
- Dirichlet PageRank (fix importance of subset)
- Weighted PageRank
- Undirected PageRank
- Timed PageRank
- PageTrust