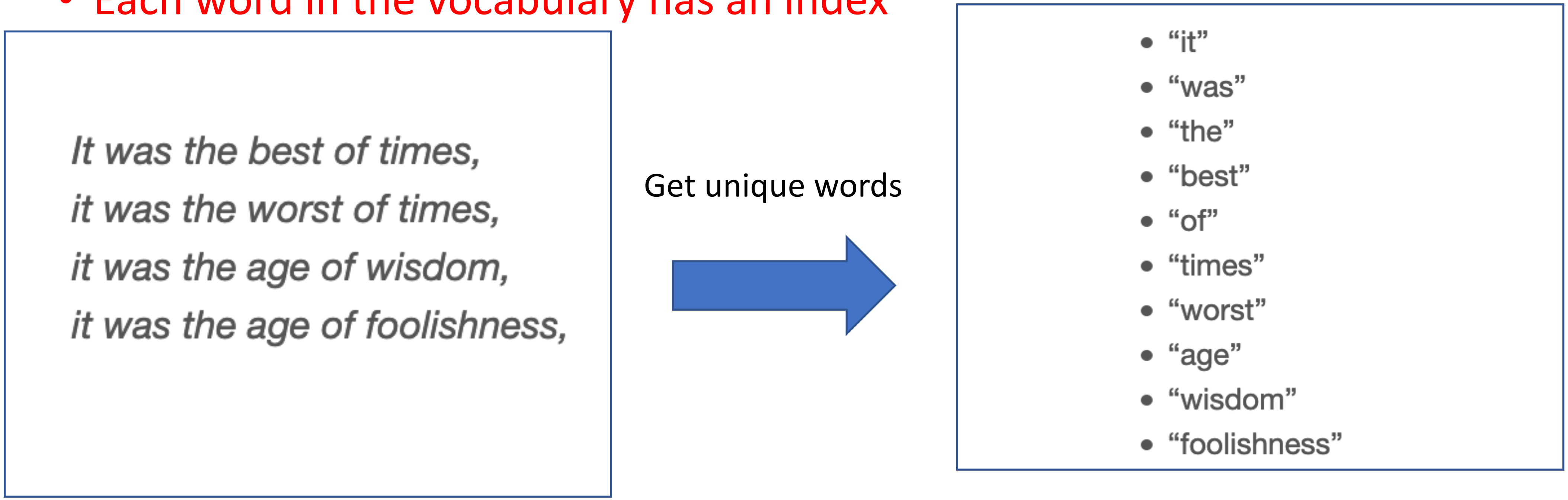
PageRank

Hongchang Gao

Spring 2022

Review of BoW

- 1. Build the vocabulary/dictionary from the given dataset
  - Get all the unique words in the given dataset
  - Each word in the vocabulary has an index



It was the best of times,
it was the worst of times,
it was the age of wisdom,
it was the age of foolishness,

Get unique words

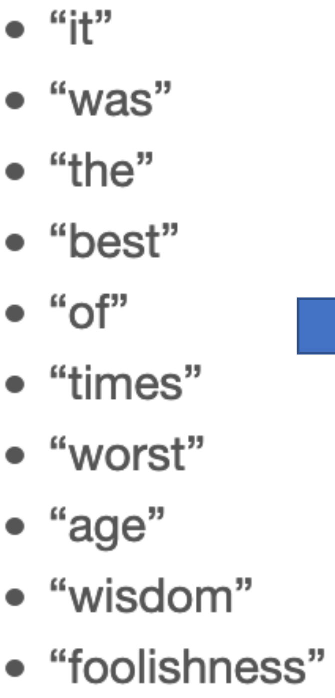
- "it"
- "was"
- "the"
- "best"
- "of"
- "times"
- "worst"
- "age"
- "wisdom"
- "foolishness"

The given dataset.
(Each sentence is a sample)

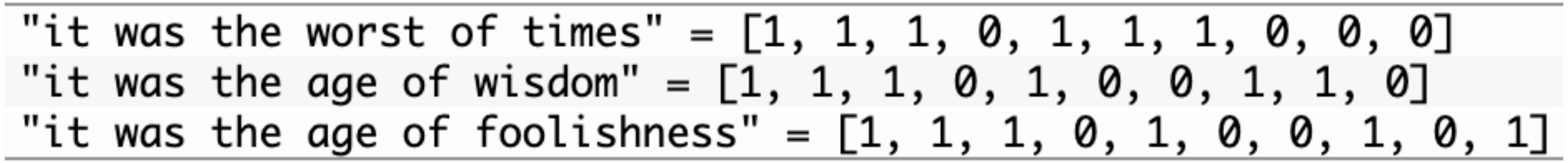
Vocabulary/dictionary
(unique words in the given dataset)

Review of BoW

- 2. Represent each sentence/paragraph/article with the vocabulary
  - Use a vector whose dimensionality equals to the size of the vocabulary
  - If the word appears, add 1 to the corresponding element in the vector



- “it”
- “was”
- “the”
- “best”
- “of”
- “times”
- “worst”
- “age”
- “wisdom”
- “foolishness”



```
"it was the worst of times" = [1, 1, 1, 0, 1, 1, 1, 0, 0, 0]
"it was the age of wisdom" = [1, 1, 1, 0, 1, 0, 0, 1, 1, 0]
"it was the age of foolishness" = [1, 1, 1, 0, 1, 0, 0, 1, 0, 1]
```

Review of BoW

- Term Frequency-Inverse Document Frequency (TF-IDF)
  - Reflect how important a word is to a document in a collection

- Definition

$$\underbrace{TF(t,d) = \frac{\#t \text{ in document } d}{\#words \text{ in document } d} \qquad\qquad IDF(t) = \log\frac{\#documents}{\#documents \text{ containing } t}}_{TF\_IDF = TF(t,d) \times IDF(t)}$$

Review of NMF

$$\min \| X - FG^T \|_F^2$$

$$s.t.\ \ F \geq 0, G \geq 0$$

- Columns of F are the underlying basis vectors

$$F = [f_1, f_2, ..., f_k]$$

- Rows of G give the weights associated with each basis vector.

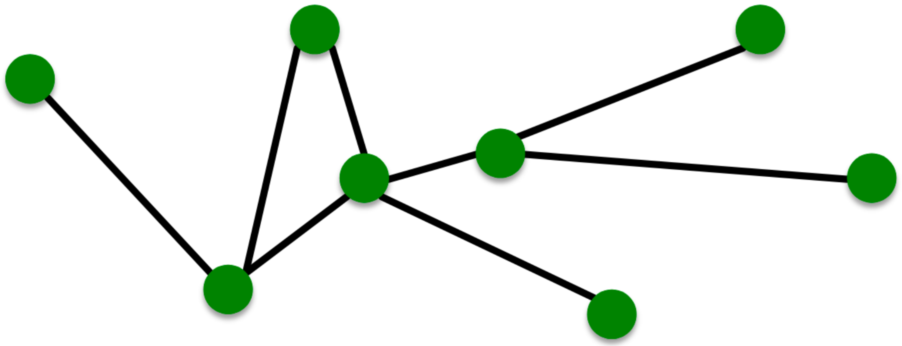
$$[\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n] = [\mathbf{f}_1, \mathbf{f}_2, \cdots, \mathbf{f}_k] \begin{bmatrix} g_{11} & g_{21} & \cdots & g_{n1} \\ g_{12} & g_{22} & \cdots & g_{n2} \\ \vdots & \vdots & \vdots & \vdots \\ g_{1k} & g_{2k} & \cdots & g_{nk} \end{bmatrix}$$

$$\mathbf{x}_i = \mathbf{f}_1 g_{i1} + \mathbf{f}_2 g_{i2} + \cdots + \mathbf{f}_k g_{ik}$$

only additive combinations!!!

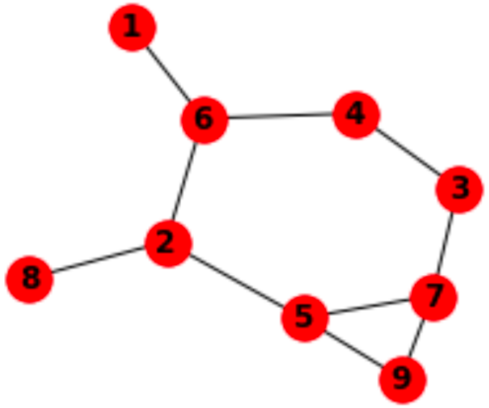
Graph Terminology

- Components of a Graph
  - Nodes/vertices
  - Edges/links
  - Graph/Network

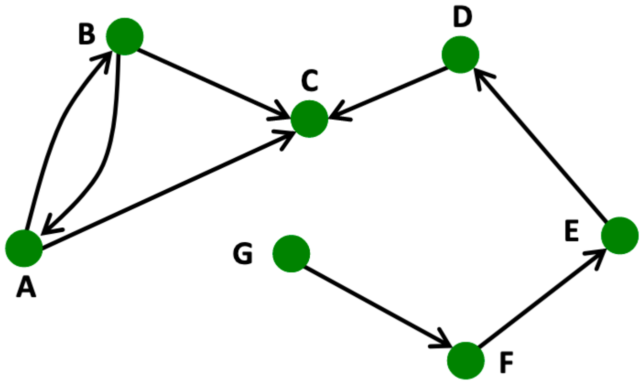


Graph Terminology

- Types of Graphs
  - Undirected Graph: links are undirected
    - Friendship on Facebook
  - Directed Graph: links are directed
    - Following on Twitter



Undirected graph

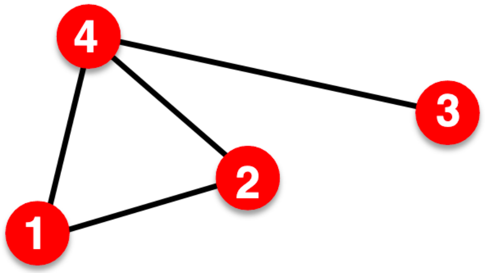


Directed graph

Graph Terminology

- Adjacency matrix

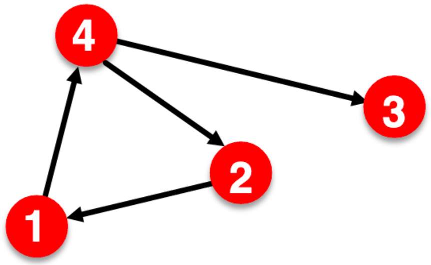
Undirected



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

$$A_{ij} = A_{ji}$$
$$A_{ii} = 0$$

Directed

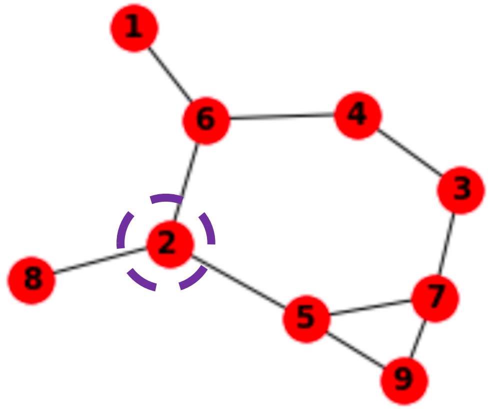


$$A = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \end{pmatrix}$$

$$A_{ij} \neq A_{ji}$$
$$A_{ii} = 0$$

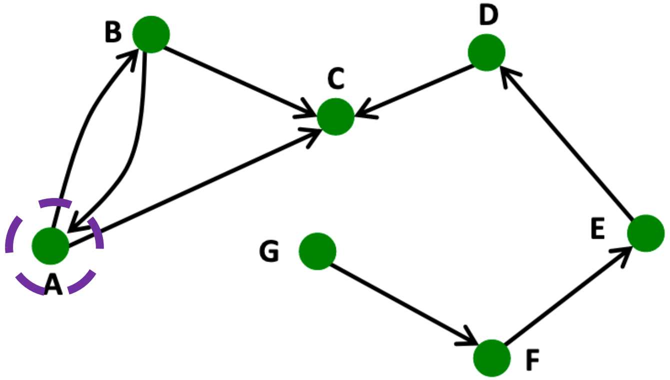
Graph Terminology

- Node degrees of undirected graph
  - The number of edges adjacent to a node



Node 2: d = 3

- Node degrees of directed graph
  - In-degree: the number of head ends adjacent to a node
  - Out-degree: the number of tail ends adjacent to a node

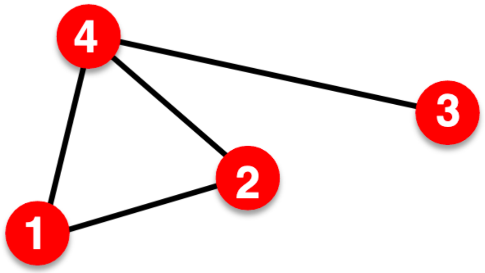


Node A: d_{in}=1, d_{out}=2

Graph Terminology

- Node degrees

Undirected

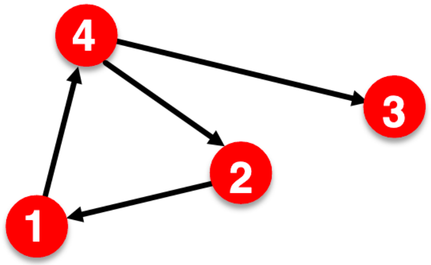


$$A = \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix}$$

$$A_{ij} = A_{ji}$$
$$A_{ii} = 0$$

Node 2: 1+0+0+1=2

Directed



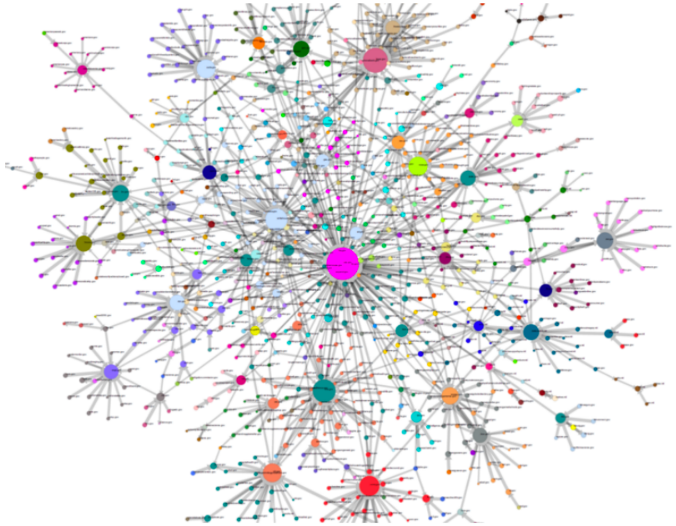
$$A = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \end{pmatrix}$$

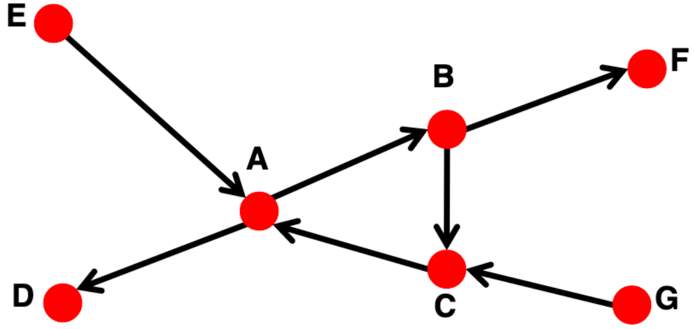
$$A_{ij} \neq A_{ji}$$
$$A_{ii} = 0$$

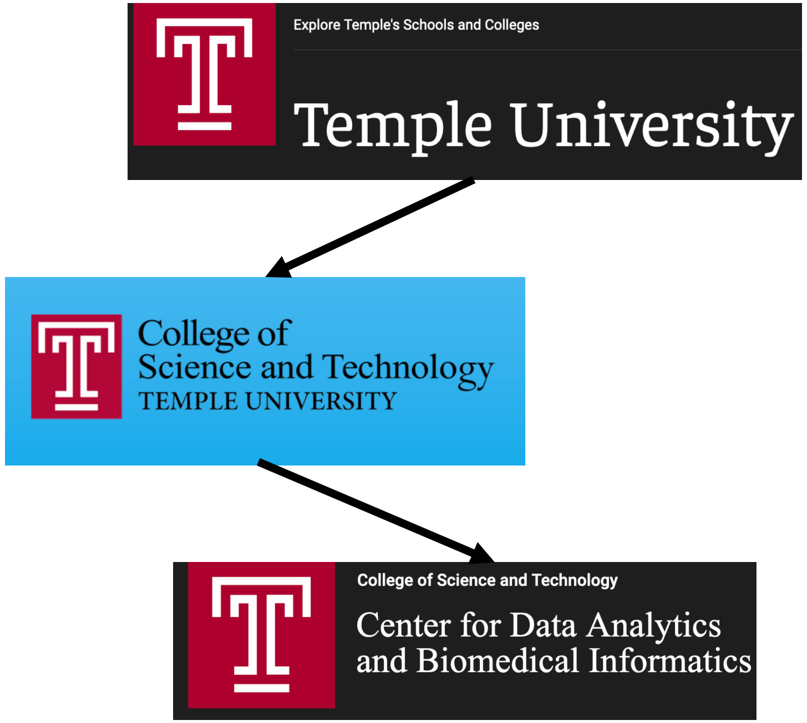
Node 4:
Indegree (column sum): 1+0+0+0=1
Outdegree (row sum): 0+1+1+0=2

PageRank

- The web is a graph
  - Nodes: web pages
  - Edges: hyperlinks







Explore Temple's Schools and Colleges

Temple University

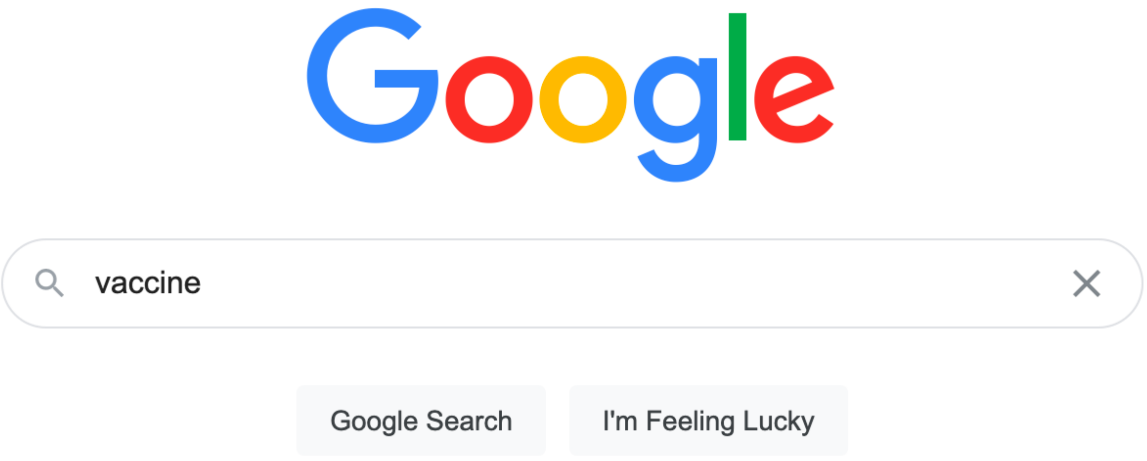
College of Science and Technology TEMPLE UNIVERSITY

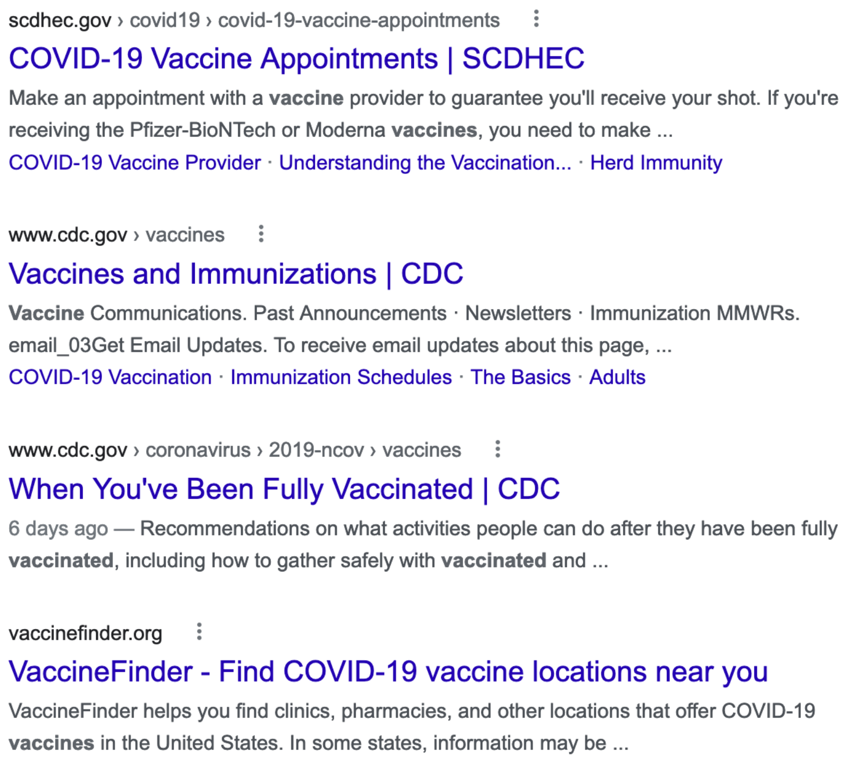
College of Science and Technology

Center for Data Analytics and Biomedical Informatics

PageRank

- All web pages are not equally “important”
  - Some webpages should be assigned more priority than others, for being more important
  - Which node (webpage) is important?





scdhec.gov › covid19 › covid-19-vaccine-appointments

COVID-19 Vaccine Appointments | SCDHEC

Make an appointment with a **vaccine** provider to guarantee you'll receive your shot. If you're receiving the Pfizer-BioNTech or Moderna **vaccines**, you need to make ...

COVID-19 Vaccine Provider · Understanding the Vaccination... · Herd Immunity

www.cdc.gov › vaccines

Vaccines and Immunizations | CDC

Vaccine Communications. Past Announcements · Newsletters · Immunization MMWRs. email_03Get Email Updates. To receive email updates about this page, ...

COVID-19 Vaccination · Immunization Schedules · The Basics · Adults

www.cdc.gov › coronavirus › 2019-ncov › vaccines

When You've Been Fully Vaccinated | CDC

6 days ago — Recommendations on what activities people can do after they have been fully **vaccinated**, including how to gather safely with **vaccinated** and ...

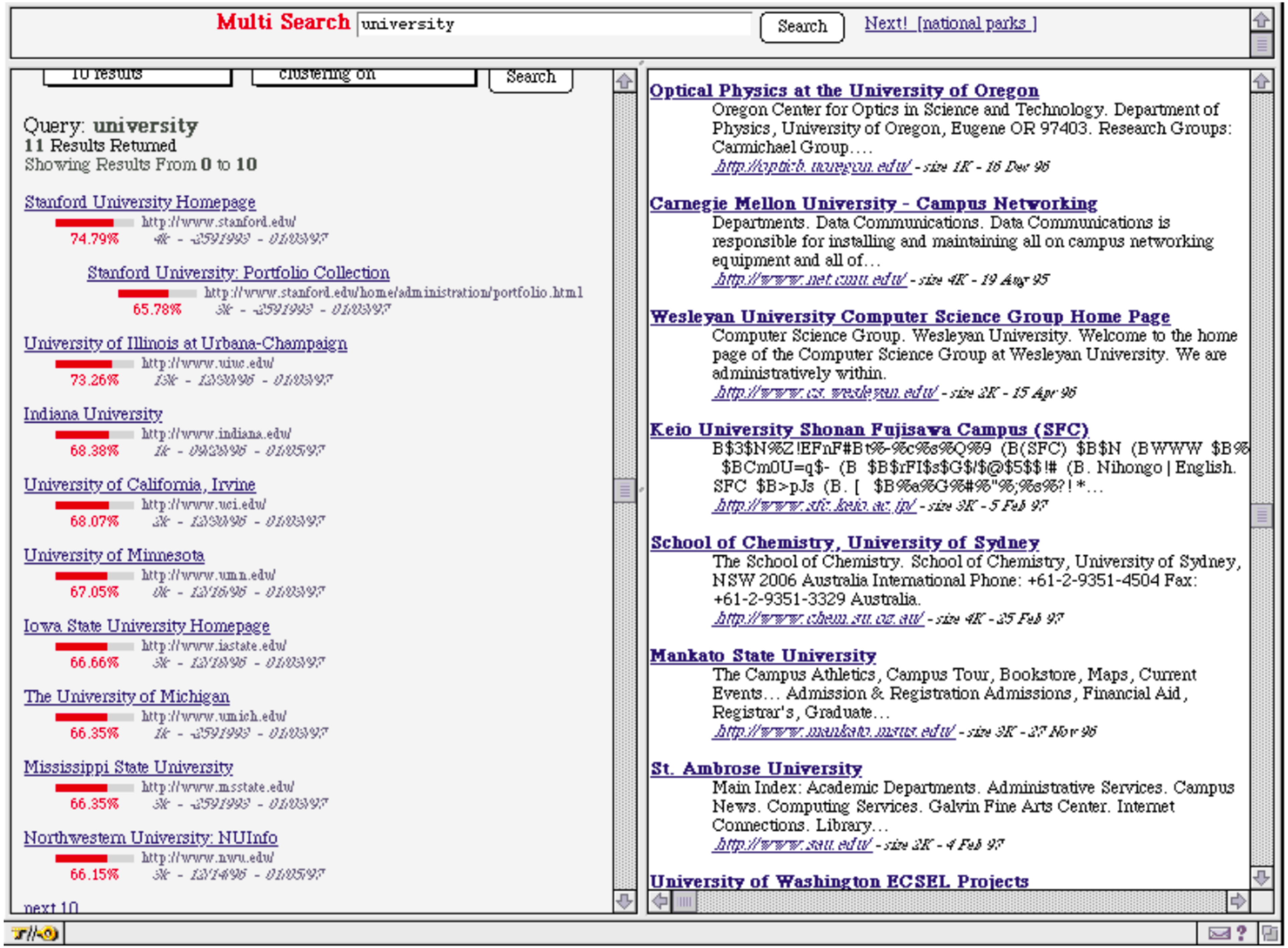
vaccinefinder.org

VaccineFinder - Find COVID-19 vaccine locations near you

VaccineFinder helps you find clinics, pharmacies, and other locations that offer COVID-19 **vaccines** in the United States. In some states, information may be ...

Pa

- Ra
  -
- Pa
  -
- Se
  - ery
  -



Query: **university**
11 Results Returned
Showing Results From **0** to **10**

Stanford University Homepage
http://www.stanford.edu/
74.79% 4K - -2591993 - 01/03/97

Stanford University: Portfolio Collection
http://www.stanford.edu/home/administration/portfolio.html
65.78% 3K - -2591993 - 01/03/97

University of Illinois at Urbana-Champaign
http://www.uiuc.edu/
73.26% 13K - 12/30/96 - 01/03/97

Indiana University
http://www.indiana.edu/
68.38% 1K - 09/28/96 - 01/05/97

University of California, Irvine
http://www.uci.edu/
68.07% 3K - 12/30/96 - 01/03/97

University of Minnesota
http://www.umn.edu/
67.05% 0K - 12/16/96 - 01/03/97

Iowa State University Homepage
http://www.iastate.edu/
66.66% 5K - 12/18/96 - 01/03/97

The University of Michigan
http://www.umich.edu/
66.35% 1K - -2591993 - 01/03/97

Mississippi State University
http://www.msstate.edu/
66.35% 5K - -2591993 - 01/03/97

Northwestern University: NUInfo
http://www.nwu.edu/
66.15% 5K - 12/14/96 - 01/05/97

next 10

Optical Physics at the University of Oregon
Oregon Center for Optics in Science and Technology. Department of Physics, University of Oregon, Eugene OR 97403. Research Groups: Carmichael Group....
http://optics.uoregon.edu/ - size 1K - 16 Dec 96

Carnegie Mellon University - Campus Networking
Departments. Data Communications. Data Communications is responsible for installing and maintaining all on campus networking equipment and all of...
http://www.net.cmu.edu/ - size 4K - 19 Aug 95

Wesleyan University Computer Science Group Home Page
Computer Science Group. Wesleyan University. Welcome to the home page of the Computer Science Group at Wesleyan University. We are administratively within.
http://www.cs.wesleyan.edu/ - size 3K - 15 Apr 96

Keio University Shonan Fujisawa Campus (SFC)
B$3$N%Z!EFnF#Bt%-%c%s%Q%9 (B(SFC) BN (BWWW $B%
$BCm0U=q$- (B BrFIsG$/$@$5$$!# (B. Nihongo | English.
SFC $B>pJs (B. [$B%a%G%#%"%;%s%?!*...
http://www.sfc.keio.ac.jp/ - size 3K - 5 Feb 97

School of Chemistry, University of Sydney
The School of Chemistry. School of Chemistry, University of Sydney, NSW 2006 Australia International Phone: +61-2-9351-4504 Fax: +61-2-9351-3329 Australia.
http://www.chem.su.oz.au/ - size 4K - 25 Feb 97

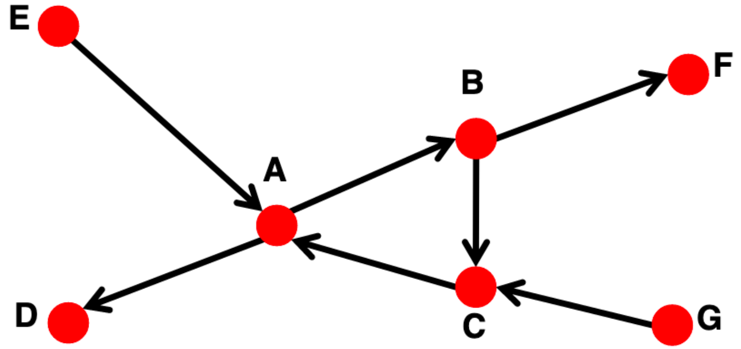
Mankato State University
The Campus Athletics, Campus Tour, Bookstore, Maps, Current Events... Admission & Registration Admissions, Financial Aid, Registrar's, Graduate...
http://www.mankato.msus.edu/ - size 3K - 27 Nov 96

St. Ambrose University
Main Index: Academic Departments. Administrative Services. Campus News. Computing Services. Galvin Fine Arts Center. Internet Connections. Library...
http://www.sau.edu/ - size 2K - 4 Feb 97

University of Washington ECSEL Projects

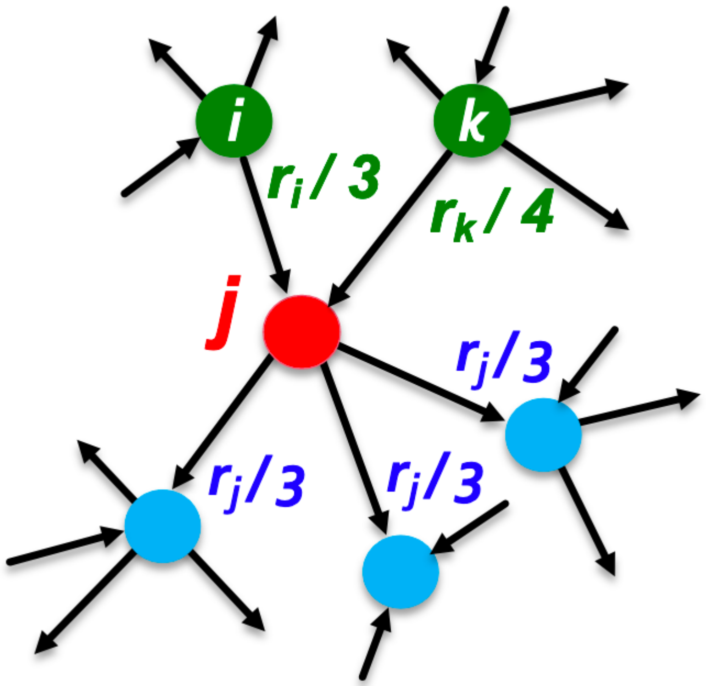
PageRank

- Compute the importance of nodes in a graph
  - Idea: Links as votes
  - Page is more important if it has more links
- Use in-links as votes
  - How to use votes to compute the importance score???
  - Q: E and C may be different. C may be more important.
  - How to differentiate their importance when they vote for A?



PageRank

- A vote from an important page is worth more:
  - Each link's vote is proportional to the **importance** of its source page
  - If page ***i*** with importance r_i has d_i out-links, each link gets r_i / d_i votes
  - Page ***j***'s own importance r_j is the sum of the votes on its in-links



$r_j = r_i/3 + r_k/4$

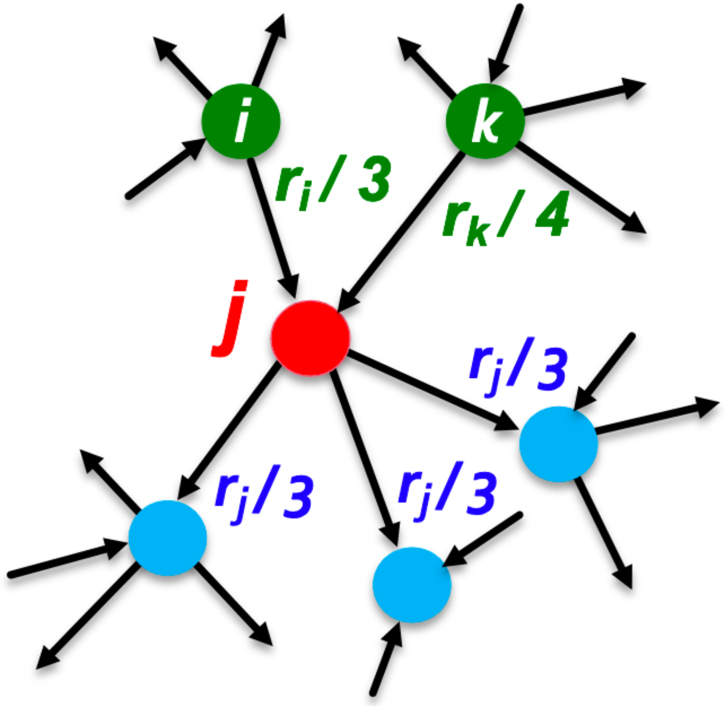
PageRank

- Formally, the importance score of each node is

$$r_j = \sum_{i \to j} \frac{r_i}{d_i}$$

d_i … out-degree of node i

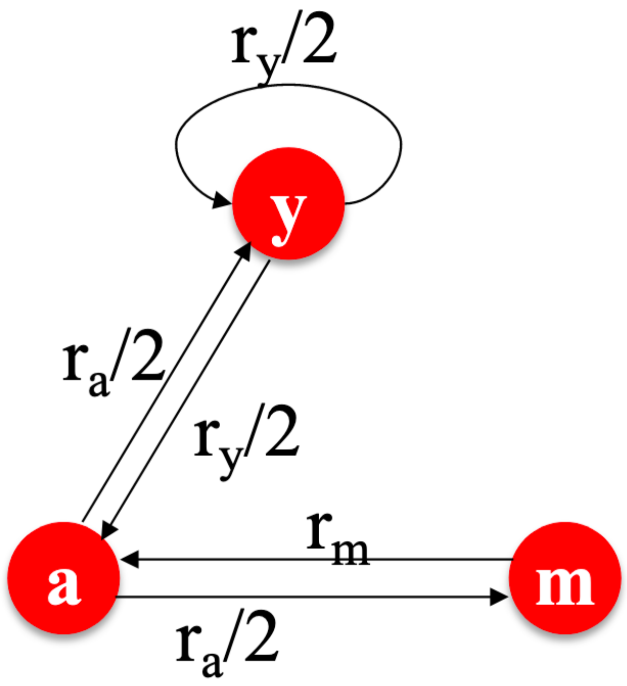
A page is important if it is pointed by other important pages



$r_j = r_i/3 + r_k/4$

PageRank

- Example



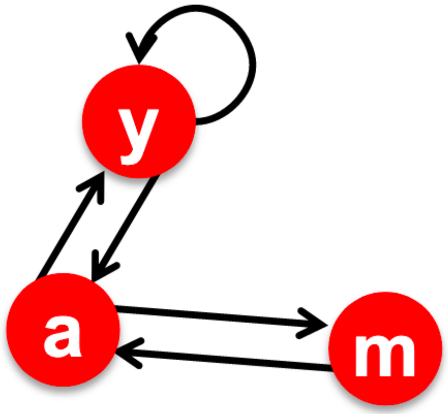
$$r_j = \sum_{i \to j} \frac{r_i}{d_i}$$

d_i … out-degree of node i

$$r_y = r_y/2 + r_a/2$$
$$r_a = r_y/2 + r_m$$
$$r_m = r_a/2$$

PageRank

- Matrix Form



Graph

Out-degree 2, 2, 1

	y	a	m
y	1	1	0
a	1	0	1
m	0	1	0

Adjacency matrix

	r_y	r_a	r_m
r_y	½	½	0
r_a	½	0	1
r_m	0	½	0

Transition matrix M

$$r_j = \sum_{i \to j} \frac{r_i}{d_i}$$

d_i … out-degree of node i

$$r_y = r_y/2 + r_a/2$$
$$r_a = r_y/2 + r_m$$
$$r_m = r_a/2$$

$$\underbrace{\begin{bmatrix} r_y \\ r_a \\ r_m \end{bmatrix}}_{r} = \underbrace{\begin{bmatrix} ½ & ½ & 0 \\ ½ & 0 & 1 \\ 0 & ½ & 0 \end{bmatrix}}_{M} \underbrace{\begin{bmatrix} r_y \\ r_a \\ r_m \end{bmatrix}}_{r}$$

PageRank

- Property of the transition matrix M
  - Column sum is 1

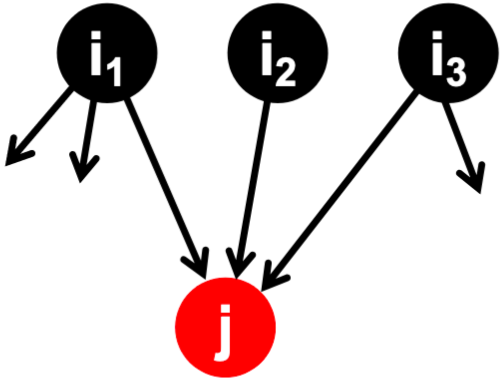
- Property of the rank vector r
  - r_i is the importance score of page i

$$\boldsymbol{\sum_i r_i \;\; = 1}$$

$$\underset{\boldsymbol{r}}{\boxed{\begin{matrix} r_y \\ r_a \\ r_m \end{matrix}}} = \underset{\boldsymbol{M}}{\boxed{\begin{matrix} \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{2} & 0 & 1 \\ 0 & \frac{1}{2} & 0 \end{matrix}}} \underset{\boldsymbol{r}}{\boxed{\begin{matrix} r_y \\ r_a \\ r_m \end{matrix}}}$$

PageRank

- Interpretation
    - At time t, the user is on page i
    - At time t+1, the user follows an out-link from i uniformly at random
    - Ends up on some page j linked from i



$$r_j = \sum_{i \to j} \frac{r_i}{d_{out}(i)}$$

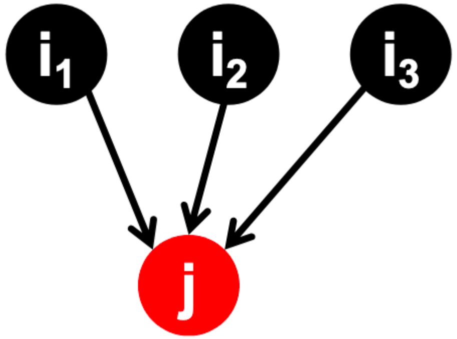
$$\underbrace{\boxed{\begin{matrix} r_y \\ r_a \\ r_m \end{matrix}}}_{r} = \underbrace{\boxed{\begin{matrix} \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{2} & 0 & 1 \\ 0 & \frac{1}{2} & 0 \end{matrix}}}_{M} \underbrace{\boxed{\begin{matrix} r_y \\ r_a \\ r_m \end{matrix}}}_{r}$$

PageRank

- Interpretation (continue):
  - Define $\boldsymbol{p(t)}$ is a probability distribution over pages
  - At time t+1, we have

$$\boldsymbol{p(t+1) = M \cdot p(t)}$$

Random walk

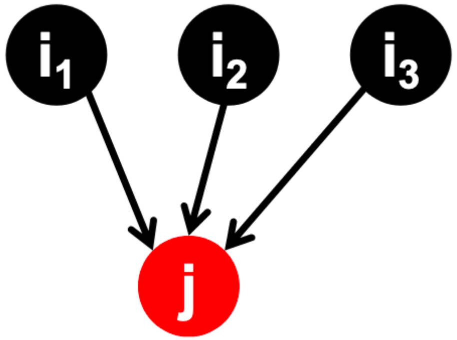


PageRank

- Solution of the importance score r:

$\boldsymbol{p}(t)$ is **stationary distribution** of a random walk

$$\boldsymbol{p}(\boldsymbol{t}+1) = \boldsymbol{M} \cdot \boldsymbol{p}(\boldsymbol{t})$$



Random walk

Stationary distribution

$$\boldsymbol{r} = \boldsymbol{M} \cdot \boldsymbol{r}$$

r is the stationary distribution of the random walk
r is the eigenvector of the transition matrix M (with eigenvalue 1)

PageRank

- Solution of the importance score r:
  - Compute the eigenvector of the transition matrix M with eigenvalue 1
  - Use power iteration to compute the eigenvector efficiently

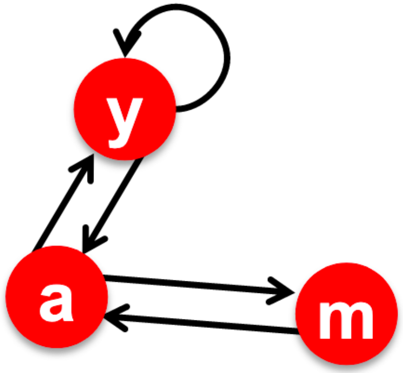
- Assign each node an initial page rank
- Repeat until convergence ($\sum_i |r_i^{t+1} - r_i^t| < \epsilon$)
  - Calculate the page rank of each node

$$r_j^{(t+1)} = \sum_{i \to j} \frac{r_i^{(t)}}{d_i}$$

- Initialize: $\boldsymbol{r}^0 = [1/N, \ldots., 1/N]^T$
- Iterate: $\boldsymbol{r}^{(\boldsymbol{t}+\boldsymbol{1})} = \boldsymbol{M} \cdot \boldsymbol{r}^t$
- Stop when $|\boldsymbol{r}^{(\boldsymbol{t}+\boldsymbol{1})} - \boldsymbol{r}^t|_1 < \varepsilon$

PageRank

- Example



	y	a	m
y	½	½	0
a	½	0	1
m	0	½	0

$$\underset{\boldsymbol{r}}{\boxed{\begin{matrix} r_y \\ r_a \\ r_m \end{matrix}}} = \underset{\boldsymbol{M}}{\boxed{\begin{matrix} ½ & ½ & 0 \\ ½ & 0 & 1 \\ 0 & ½ & 0 \end{matrix}}} \underset{\boldsymbol{r}}{\boxed{\begin{matrix} r_y \\ r_a \\ r_m \end{matrix}}}$$

$$\begin{pmatrix} r_y \\ r_a \\ r_m \end{pmatrix} = \begin{pmatrix} 1/3 \\ 1/3 \\ 1/3 \end{pmatrix} \begin{pmatrix} 1/3 \\ 3/6 \\ 1/6 \end{pmatrix} \begin{pmatrix} 5/12 \\ 1/3 \\ 3/12 \end{pmatrix} \begin{pmatrix} 9/24 \\ 11/24 \\ 1/6 \end{pmatrix} \ldots \begin{pmatrix} 6/15 \\ 6/15 \\ 3/15 \end{pmatrix}$$

Iteration 0, 1, 2, …