GOETHE UNIVERSITÄT FRANKFURT AM MAIN

www.uni-frankfurt.de

SoSe 2014: M-TANI: Big Data Analytics

Lecture 7 – 18/06/2014

Sead Izberovic

Dr. Nikolaos Korfiatis

Agenda

- **Recap from the previous session**
- **Topic specific PageRank**
- **TrustRank** (Stanford slides)
  - Link Spam (Stanford slides)
- **Hypertext-Induced Topic Selection** (Stanford slides)
  - Hubs and Authorities (Stanford slides)

PageRank

- **Principle of votes**
  - The importance r_j of page j is the sum of the votes on its in-links
  - The weight of each link is $\frac{r_j}{n}$, with n the sum of out-links from the page j
  - The ***rank*** for page j is defined by: $r_j = \sum_{i \to j} \frac{r_i}{d_i}$
    - d_i is the out-degree of the page i
  - **A vote from a *important* page is more worth then a vote from a *not-important* page**

from [2]

PageRank

- **The flow equations $r_j = \sum_{i \to j} \frac{r_i}{d_i}$ can be rewritten as $r = M \cdot r$**
  - The **rank vector r** is an **eigenvector** of the stochastic web matrix M
    - M is a **column stochastic matrix** → **The c**olumns sum to 1
- **We can now efficiently solve for r with the Power iteration method**

from [2]

Power Iteration Method

- **Power Iteration**
  - Suppose there are $\boldsymbol{N}$ web pages
  - Initialize: $r_0 = \begin{bmatrix} \frac{1}{N} \\ \vdots \\ \frac{1}{N} \end{bmatrix}$
  - Iterate: $r_{t+1} = \boldsymbol{M} \cdot r_t$
  - **Stop when** $|r_{t+1} - r_t| < \varepsilon$

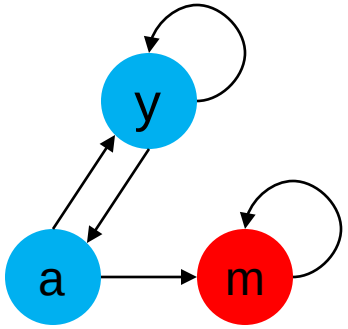
from [2]

PageRank Problems: Spider Traps

- **Power Iteration**
  - **Set $r_j = 1$**
  - $r_j = \sum_{i \to j} \frac{r_i}{d_i}$
  - d_i is the out-degree of the page i

	y	a	m
y	$\frac{1}{2}$	$\frac{1}{2}$	0
a	$\frac{1}{2}$	0	0
m	0	0	1

```
Iteration 1        Matrix M:              r0:
[[ 0.33333333]     [[ 0.5  0.5  0. ]      [[ 0.33333333]
 [ 0.16666667]  =   [ 0.5  0.   0. ]   .    [ 0.33333333]
 [ 0.5       ]]     [ 0.   0.5  1. ]]      [ 0.33333333]]
```



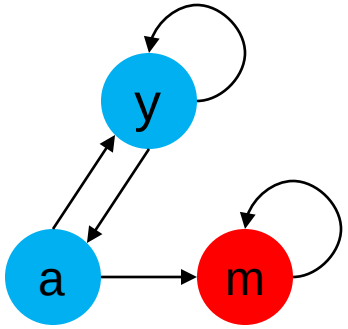
from [2]

PageRank Problems: Spider Traps

- **Power Iteration**
  - **Set $r_j = 1$**
  - $r_j = \sum_{i \to j} \frac{r_i}{d_i}$
  - d_i is the out-degree of the page i

	y	a	m
y	$\frac{1}{2}$	$\frac{1}{2}$	0
a	$\frac{1}{2}$	0	0
m	0	0	1

```
Iteration 2          Matrix M:                Iteration 1
[[ 0.25      ]       [[ 0.5  0.5  0. ]        [[ 0.33333333]
 [ 0.16666667]   =    [ 0.5  0.   0. ]    .    [ 0.16666667]
 [ 0.58333333]]       [ 0.   0.5  1. ]]       [ 0.5       ]]
```



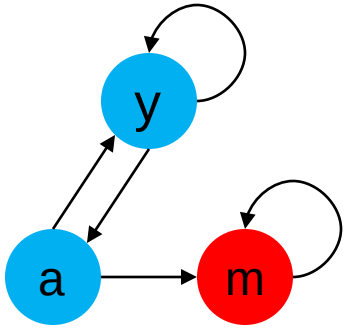
from [2]

PageRank Problems: Spider Traps

- **Power Iteration**
  - **Set $r_j = 1$**
  - $r_j = \sum_{i \to j} \frac{r_i}{d_i}$
  - d_i is the out-degree of the page i

	y	a	m
y	$\frac{1}{2}$	$\frac{1}{2}$	0
a	$\frac{1}{2}$	0	0
m	0	0	1

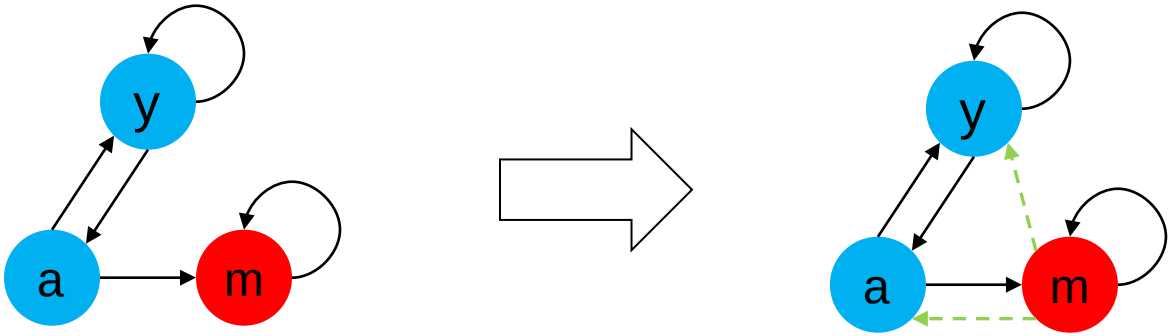
```
Iteration 20      Matrix M:                  Iteration 19
[[ 0.00563018]    [[ 0.5  0.5  0. ]     [[ 0.00695928]
 [ 0.00347964]  =  [ 0.5  0.   0. ]  .   [ 0.00430107]
 [ 0.99089019]]    [ 0.   0.5  1. ]]     [ 0.98873965]]
```



from [2]

Spider Traps Solution

- **Teleports**
  - With prob. $\boldsymbol{\beta}$, follow a link at random
  - With prob. $\mathbf{1} - \boldsymbol{\beta}$, jump to some random page



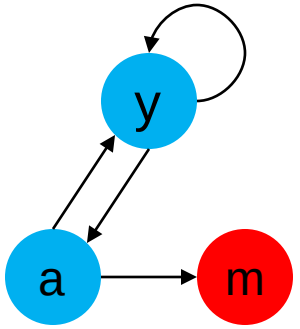
from [2]

PageRank Problems: Dead Ends

- **Power Iteration**
  - **Set $r_j = 1$**
  - $r_j = \sum_{i \to j} \frac{r_i}{d_i}$
  - d_i is the out-degree of the page i

	y	a	m
y	$\frac{1}{2}$	$\frac{1}{2}$	0
a	$\frac{1}{2}$	0	0
m	0	$\frac{1}{2}$	0

```
Iteration 1        Matrix M:                r0:
[[ 0.33333333]     [[ 0.5  0.5  0. ]        [[ 0.33333333]
 [ 0.16666667]  =   [ 0.5  0.   0. ]     .    [ 0.33333333]
 [ 0.16666667]]     [ 0.   0.5  0. ]]        [ 0.33333333]]
```



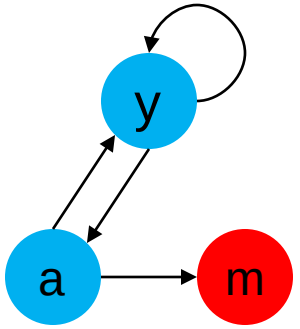
from [2]

PageRank Problems: Dead Ends

- **Power Iteration**
  - **Set $r_j = 1$**
  - $r_j = \sum_{i \to j} \frac{r_i}{d_i}$
  - d_i is the out-degree of the page i

	y	a	m
y	$\frac{1}{2}$	$\frac{1}{2}$	0
a	$\frac{1}{2}$	0	0
m	0	$\frac{1}{2}$	0

```
Iteration 2          Matrix M:              Iteration 1
[[ 0.25      ]   =   [[ 0.5  0.5  0. ]   .   [[ 0.33333333]
 [ 0.16666667]       [ 0.5  0.   0. ]       [ 0.16666667]
 [ 0.08333333]]      [ 0.   0.5  0. ]]      [ 0.16666667]]
```



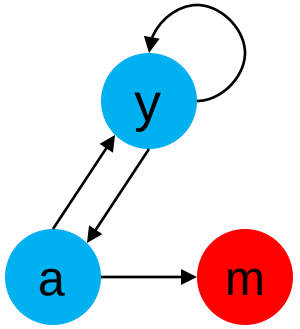
from [2]

PageRank Problems: Dead Ends

- **Power Iteration**
  - **Set $r_j = 1$**
  - $r_j = \sum_{i \to j} \frac{r_i}{d_i}$
  - d_i is the out-degree of the page i

	y	a	m
y	$\frac{1}{2}$	$\frac{1}{2}$	0
a	$\frac{1}{2}$	0	0
m	0	$\frac{1}{2}$	0

```
Iteration 20          Matrix M:                Iteration 19
[[ 0.00563018]   =    [[ 0.5  0.5  0. ]    .   [[ 0.00695928]
 [ 0.00347964]         [ 0.5  0.   0. ]         [ 0.00430107]
 [ 0.00215054]]        [ 0.   0.5  0. ]]        [ 0.00265821]]
```



from [2]

PageRank Problems: Dead Ends

- **Power Iteration**
  - **Set $r_j = 1$**
  - $r_j = \sum_{i \to j} \frac{r_i}{d_i}$
  - d_i is the out-degree of the page i

	y	a	m
y	$\frac{1}{2}$	$\frac{1}{2}$	0
a	$\frac{1}{2}$	0	0
m	0	$\frac{1}{2}$	0

```
Iteration 20
[[ 0.00563018]
 [ 0.00347964]
 [ 0.00215054]]
```

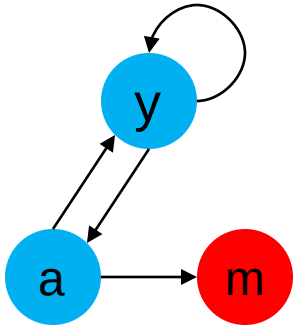
=

```
Matrix M:
[[ 0.5  0.5  0. ]
 [ 0.5  0.   0. ]
 [ 0.   0.5  0. ]]
```

·

```
Iteration 19
[[ 0.00695928]
 [ 0.00430107]
 [ 0.00265821]]
```

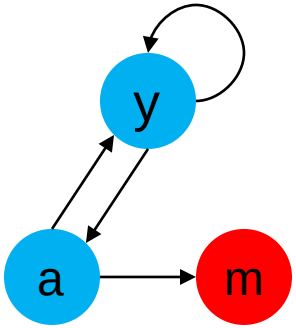
Matrix is **not** column stochastic

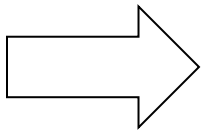


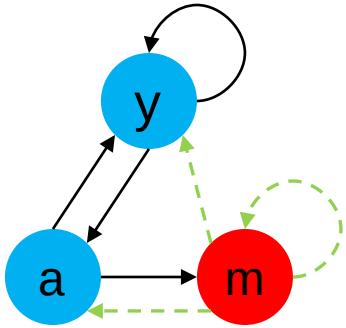
from [2]

Dead Ends Solution

- **Teleports**
  - Follow random teleport links with probability 1.0 from dead-ends







	y	a	m
y	$\frac{1}{2}$	$\frac{1}{2}$	0
a	$\frac{1}{2}$	0	0
m	0	$\frac{1}{2}$	0

	y	a	m
y	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{3}$
a	$\frac{1}{2}$	0	$\frac{1}{3}$
m	0	$\frac{1}{2}$	$\frac{1}{3}$

from [2]

Google Matrix

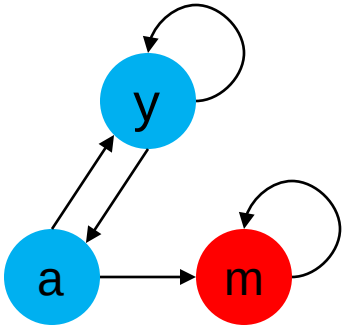
- **PageRank equation** $r_j = \sum_{i \to j} \beta \frac{r_i}{d_i} + (1 - \beta) \frac{1}{N}$
  - With prob. β, follow a link at random
  - With prob. $1 - \beta$, jump to some random page

- **Google Matrix** A **:**
  - $A = \beta M + (1 - \beta) \left[\frac{1}{N}\right]_{N \times N}$
  - $r = A \cdot r \rightarrow$ Power Iteration works

All entries are $\frac{1}{N}$

from [2]

Google Matrix Example



$\beta = 0.8$

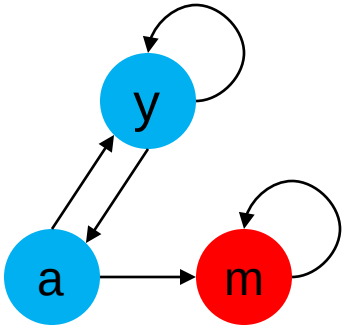
	y	a	m
y	$\frac{1}{2}$	$\frac{1}{2}$	0
a	$\frac{1}{2}$	0	0
m	0	0	1

$$A = 0.8 \cdot \begin{matrix}\text{Matrix M:}\\ \begin{bmatrix} 0.5 & 0.5 & 0. \\ 0.5 & 0. & 0. \\ 0. & 0.5 & 1. \end{bmatrix}\end{matrix} + 0.2 \cdot \begin{matrix}\text{1/n}\\ \begin{bmatrix} 0.33333333 & 0.33333333 & 0.33333333 \\ 0.33333333 & 0.33333333 & 0.33333333 \\ 0.33333333 & 0.33333333 & 0.33333333 \end{bmatrix}\end{matrix}$$

$$A = \begin{bmatrix} 0.46666667 & 0.46666667 & 0.06666667 \\ 0.46666667 & 0.06666667 & 0.06666667 \\ 0.06666667 & 0.46666667 & 0.86666667 \end{bmatrix}$$

from [2]

Google Matrix Example



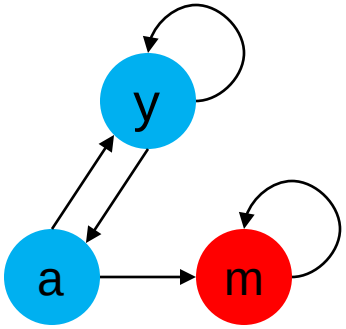
$$A = \begin{bmatrix} 0.46666667 & 0.46666667 & 0.06666667 \\ 0.46666667 & 0.06666667 & 0.06666667 \\ 0.06666667 & 0.46666667 & 0.86666667 \end{bmatrix}$$

Power Iteration

$$\text{Iteration 1} \begin{bmatrix} 0.33333333 \\ 0.2 \\ 0.46666667 \end{bmatrix} = \begin{bmatrix} 0.46666667 & 0.46666667 & 0.06666667 \\ 0.46666667 & 0.06666667 & 0.06666667 \\ 0.06666667 & 0.46666667 & 0.86666667 \end{bmatrix} \cdot \begin{matrix} \text{r0:} \\ \begin{bmatrix} 0.33333333 \\ 0.33333333 \\ 0.33333333 \end{bmatrix} \end{matrix}$$

from [2]

Google Matrix Example



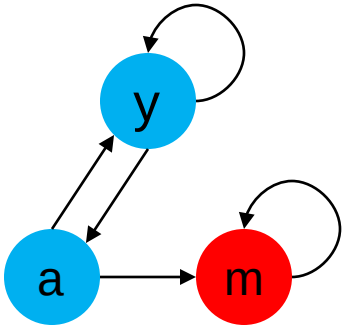
$$A = \begin{bmatrix} 0.46666667 & 0.46666667 & 0.06666667 \\ 0.46666667 & 0.06666667 & 0.06666667 \\ 0.06666667 & 0.46666667 & 0.86666667 \end{bmatrix}$$

Power Iteration

$$\underset{\text{Iteration 2}}{\begin{bmatrix} 0.28 \\ 0.2 \\ 0.52 \end{bmatrix}} = \begin{bmatrix} 0.46666667 & 0.46666667 & 0.06666667 \\ 0.46666667 & 0.06666667 & 0.06666667 \\ 0.06666667 & 0.46666667 & 0.86666667 \end{bmatrix} \cdot \underset{\text{Iteration 1}}{\begin{bmatrix} 0.33333333 \\ 0.2 \\ 0.46666667 \end{bmatrix}}$$

from [2]

Google Matrix Example



$$A = \begin{bmatrix} 0.46666667 & 0.46666667 & 0.06666667 \\ 0.46666667 & 0.06666667 & 0.06666667 \\ 0.06666667 & 0.46666667 & 0.86666667 \end{bmatrix}$$

Power Iteration

$$\text{Iteration 20} \begin{bmatrix} 0.21214932 \\ 0.15153253 \\ 0.63631815 \end{bmatrix} = \begin{bmatrix} 0.46666667 & 0.46666667 & 0.06666667 \\ 0.46666667 & 0.06666667 & 0.06666667 \\ 0.06666667 & 0.46666667 & 0.86666667 \end{bmatrix} \cdot \text{Iteration 19} \begin{bmatrix} 0.21216465 \\ 0.151542 \\ 0.63629336 \end{bmatrix}$$

from [2]

PageRank Problems

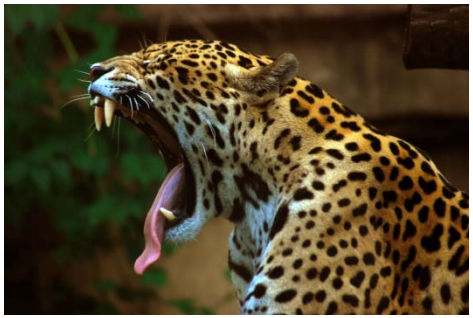
- **Measures generic popularity of a page**
  - Ignores topic-specific authorities
  - **Solution**: Topic-Specific/Sensitive PageRank

from [2]

Topic-Specific PageRank

- **Goal:** Evaluate Web pages not just according to their popularity, but by how close they are to a particular topic, e.g. “sports” or “history”
- Allows search queries to be answered based on interests of the user

Example: Search query = **`jaguar`**



X

from [1] and [2]

Topic-Specific PageRank

- **Idea:** biasing the PageRank to favor pages that share same topic
- Difference to the standard PageRank
  - Standard PageRank
    - **Teleport can go to any page with equal probability**
  - Topic Specific PageRank
    - **Teleport can go to a topic-specific set of "relevant" pages (teleport set)**

from [2]

Topic-Specific PageRank

- **What is the teleport set S?**
  - S contains only pages that are relevant to a specific topic
- **What are the benefits of using the teleport set?**
  - For each teleport set S, we get a different (*topic specific*) rank vector r_S

from [2]

Topic-Specific PageRank

- **Matrix formulation**
  - Standard PageRank

$$\boldsymbol{A = \beta M + (1 - \beta)\left[\frac{1}{N}\right]_{N \times N}}$$

  - Topic-Specific PageRank

$$\boldsymbol{A_{ij} = \begin{cases} \beta M_{ij} + \dfrac{(1-\beta)}{|S|} \; if \; i \in S \\ \beta M_{ij} + 0 \; if \; i \notin S \end{cases}}$$

from [2]

Topic-Specific PageRank

- **Matrix formulation**
  - Vector $\boldsymbol{e_S}$

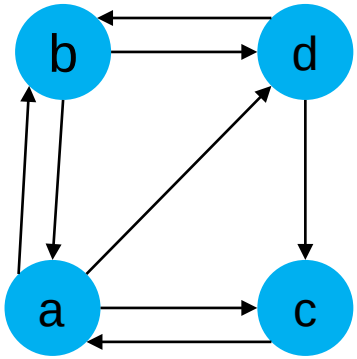
$$e_{S_i} = \begin{cases} \frac{(1-\beta)}{|S|} \; if \; i \in S \\ 0 \; if \; i \notin S \end{cases}$$

  - Topic-Specific PageRank

$$A = \beta M + e_S$$

from [1] and [2]

Topic-Specific PageRank Example



$\boldsymbol{\beta = 0.8; S = \{b, d\}}$

```
Matrix M:
[[ 0.        0.5       1.        0.      ]
 [ 0.333333  0.        0.        0.5     ]
 [ 0.333333  0.        0.        0.5     ]
 [ 0.333333  0.5       0.        0.      ]]
```

$\boldsymbol{\beta M} =$

```
[[ 0.        0.4       0.8       0.      ]
 [ 0.2666664 0.        0.        0.4     ]
 [ 0.2666664 0.        0.        0.4     ]
 [ 0.2666664 0.4       0.        0.      ]]
```

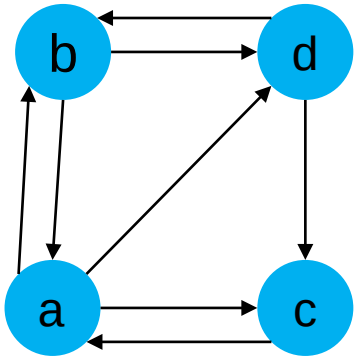
← Not stochastic

$\boldsymbol{e_S} =$

```
[[ 0. ]
 [ 0.1]
 [ 0. ]
 [ 0.1]]
```

from [1]

Topic-Specific PageRank Example



$\boldsymbol{\beta = 0.8; S = \{b, d\}}$

```
Matrix M:
[[ 0.        0.5       1.        0.      ]
 [ 0.333333  0.        0.        0.5     ]
 [ 0.333333  0.        0.        0.5     ]
 [ 0.333333  0.5       0.        0.      ]]
```

$A =$

```
[[ 0.        0.4       0.8       0.      ]
 [ 0.2666664 0.        0.        0.4     ]
 [ 0.2666664 0.        0.        0.4     ]
 [ 0.2666664 0.4       0.        0.      ]]
```

$+$

```
[[ 0. ]
 [ 0.1]
 [ 0. ]
 [ 0.1]]
```

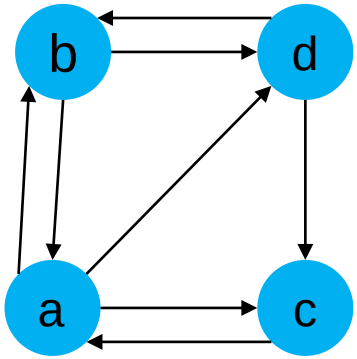
$=$

```
[[ 0.        0.4       0.8       0.      ]
 [ 0.3666664 0.1       0.1       0.5     ]
 [ 0.2666664 0.        0.        0.4     ]
 [ 0.3666664 0.5       0.1       0.1     ]]
```

← stochastic!

from [1]

Topic-Specific PageRank Example



$\boldsymbol{\beta = 0.8; S = \{b, d\}}$

```
Matrix M:
[[ 0.        0.5       1.        0.      ]
 [ 0.333333  0.        0.        0.5     ]
 [ 0.333333  0.        0.        0.5     ]
 [ 0.333333  0.5       0.        0.      ]]
```

$A =$

```
[[ 0.        0.4       0.8       0.      ]
 [ 0.2666664 0.        0.        0.4     ]
 [ 0.2666664 0.        0.        0.4     ]
 [ 0.2666664 0.4       0.        0.      ]]
```

$+$

```
[[ 0. ]
 [ 0.1]
 [ 0. ]
 [ 0.1]]
```

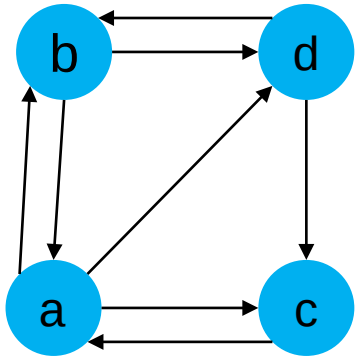
$=$

```
[[ 0.        0.4       0.8       0.      ]
 [ 0.3666664 0.1       0.1       0.5     ]
 [ 0.2666664 0.        0.        0.4     ]
 [ 0.3666664 0.5       0.1       0.1     ]]
```

← stochastic!

from [1]

Topic-Specific PageRank Example



$\beta = 0.8; S = \{b, d\}$

```
Matrix M:
[[ 0.        0.5       1.        0.      ]
 [ 0.333333  0.        0.        0.5     ]
 [ 0.333333  0.        0.        0.5     ]
 [ 0.333333  0.5       0.        0.      ]]
```

Power Iteration

```
Iteration 1
[[ 0.3      ]
 [ 0.2666666]
 [ 0.1666666]
 [ 0.2666666]]
```

=

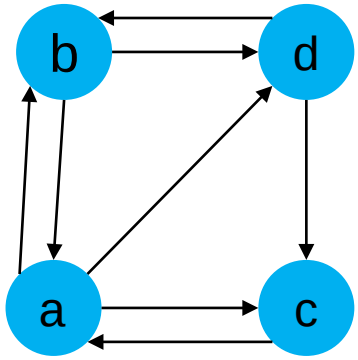
```
[[ 0.        0.4       0.8       0.      ]
 [ 0.3666664 0.1       0.1       0.5     ]
 [ 0.2666664 0.        0.        0.4     ]
 [ 0.3666664 0.5       0.1       0.1     ]]
```

·

```
r0:
[[ 0.25]
 [ 0.25]
 [ 0.25]
 [ 0.25]]
```

from [1]

Topic-Specific PageRank Example



$\beta = 0.8; S = \{b, d\}$

```
Matrix M:
[[ 0.        0.5       1.        0.      ]
 [ 0.333333  0.        0.        0.5     ]
 [ 0.333333  0.        0.        0.5     ]
 [ 0.333333  0.5       0.        0.      ]]
```

Power Iteration

```
Iteration 2
[[ 0.23999992]
 [ 0.28666654]
 [ 0.18666656]
 [ 0.28666654]]
```

$=$

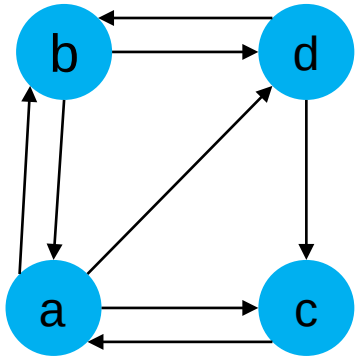
```
[[ 0.        0.4       0.8       0.      ]
 [ 0.3666664 0.1       0.1       0.5     ]
 [ 0.2666664 0.        0.        0.4     ]
 [ 0.3666664 0.5       0.1       0.1     ]]
```

$\cdot$

```
Iteration 1
[[ 0.3      ]
 [ 0.2666666]
 [ 0.1666666]
 [ 0.2666666]]
```

from [1]

Topic-Specific PageRank Example



$\beta = 0.8; S = \{b, d\}$

```
Matrix M:
[[ 0.        0.5       1.        0.      ]
 [ 0.333333  0.        0.        0.5     ]
 [ 0.333333  0.        0.        0.5     ]
 [ 0.333333  0.5       0.        0.      ]]
```

Power Iteration

```
Iteration 20
[[ 0.25714183]
 [ 0.28095122]
 [ 0.18095161]
 [ 0.28095122]]
```

=

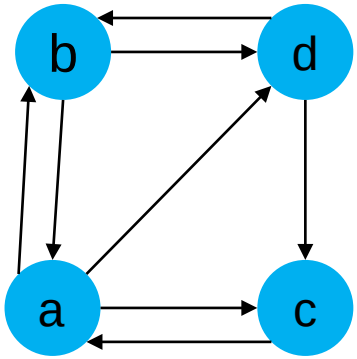
```
[[ 0.        0.4       0.8       0.      ]
 [ 0.3666664 0.1       0.1       0.5     ]
 [ 0.2666664 0.        0.        0.4     ]
 [ 0.3666664 0.5       0.1       0.1     ]]
```

·

```
Iteration 19
[[ 0.25714188]
 [ 0.28095127]
 [ 0.18095164]
 [ 0.28095127]]
```

from [1]

Topic-Specific PageRank Example



$\boldsymbol{\beta = 0.8; S = \{b, d\}}$

```
Matrix M:
[[ 0.        0.5       1.        0.      ]
 [ 0.333333  0.        0.        0.5     ]
 [ 0.333333  0.        0.        0.5     ]
 [ 0.333333  0.5       0.        0.      ]]
```

Topic-Specific PageRank

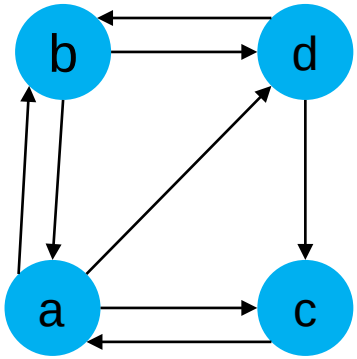
```
Iteration 20
[[ 0.25714183]
 [ 0.28095122]
 [ 0.18095161]
 [ 0.28095122]]
```

Standard PageRank

```
Iteration 20
[[ 0.33333112]
 [ 0.22222075]
 [ 0.22222075]
 [ 0.22222075]]
```

from [1] and [2]

Topic-Specific PageRank Example



$\boldsymbol{\beta = 0.8; S = \{b, d\}}$

```
Matrix M:
[[ 0.        0.5       1.        0.      ]
 [ 0.333333  0.        0.        0.5     ]
 [ 0.333333  0.        0.        0.5     ]
 [ 0.333333  0.5       0.        0.      ]]
```

Topic-Specific PageRank

```
Iteration 20
[[ 0.25714183]
 [ 0.28095122]
 [ 0.18095161]
 [ 0.28095122]]
```

Standard PageRank

```
Iteration 20
[[ 0.33333112]
 [ 0.22222075]
 [ 0.22222075]
 [ 0.22222075]]
```

from [1] and [2]

Topic-Specific PageRank

- Who to create the teleport set S?
  - User can pick the topic from a menu
  - Classify query into a topic
  - Using context of the query
    - History of queries e.g. “video games” followed by “jaguar”
  - Using user context
    - Bookmarks
    - Browser History....

from [2]

Literature

1. Anand Rajaraman, Jeffrey D. Ullman, Jure Leskovec. 2014
   Mining of Massive Datasets
   Cambridge University Press
2. Jure Leskovec. 2014
   Slides: **Mining Massive Data Sets**
   URL: http://www.stanford.edu/class/cs246/slides/09-pagerank.pdf