CSE 4/587
Data Intensive Computing

Dr. Eric Mikida
epmikida@buffalo.edu
208 Capen Hall

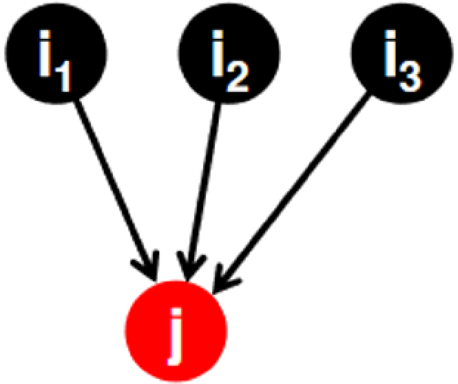
Dr. Shamshad Parvin
shamsadp@buffalo.edu
313 Davis Hall

Graph Analysis and Page Rank

Review from previous Lecture

- We have looked the Page Rank in the form of
- ---Flow Model Equations
- -- Matrix Formulations
- Find out the Eigen vector form the Matrix Formulations for the Page rank

Random Walk Interpretation

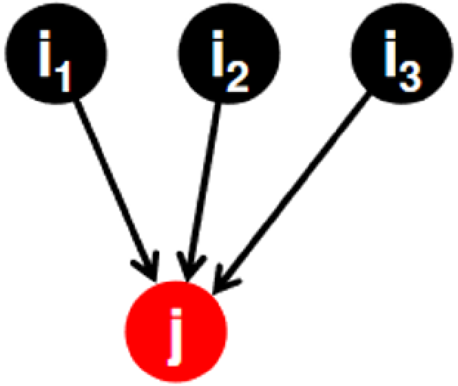


$$r_j = \sum_{i \to j} \frac{r_i}{\mathrm{d}_{\mathrm{out}}(\mathrm{i})}$$

Random Walk Interpretation

Imagine a random web surfer

- At any time ***t***, the surfer is on some page ***i***

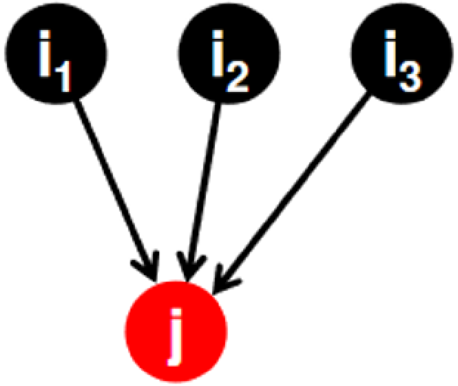


$$r_j = \sum_{i \to j} \frac{r_i}{\mathrm{d_{out}(i)}}$$

Random Walk Interpretation

Imagine a random web surfer

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

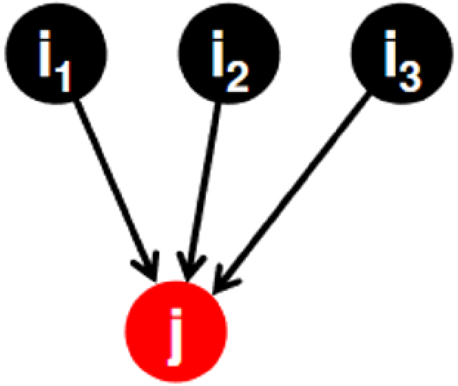


$$r_j = \sum_{i \to j} \frac{r_i}{\mathrm{d}_{\mathrm{out}}(\mathrm{i})}$$

Random Walk Interpretation

Imagine a random web surfer

- At any time ***t***, the surfer is on some page ***i***
- At time ***t* + 1**, the surfer follows an out-link from ***i*** uniformly at random
    - Ends up on some page ***j*** linked from ***i***
- Process repeats infinitely



$$r_j = \sum_{i \to j} \frac{r_i}{\mathrm{d}_{\mathrm{out}}(\mathrm{i})}$$

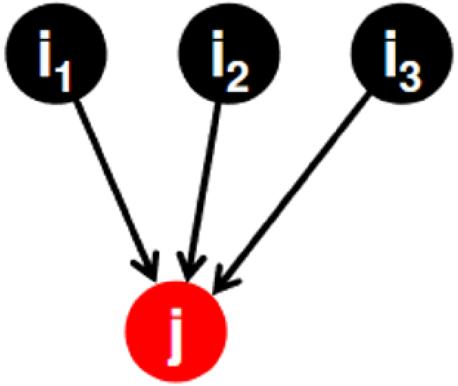
Random Walk Interpretation

Imagine a random web surfer

- At any time ***t***, the surfer is on some page ***i***
- At time ***t* + 1**, the surfer follows an out-link from ***i*** uniformly at random
    - Ends up on some page ***j*** linked from ***i***
- Process repeats infinitely

P*(*t*)** is the vector whose ***ith coordinate is the probability that the surfer is at page ***i*** at time ***t***

So ***P*(*t*)** is a probability distribution over pages



$$r_j = \sum_{i \to j} \frac{r_i}{\mathrm{d}_{\mathrm{out}}(\mathrm{i})}$$

Random Walk Interpretation

- **Where is the surfer at time *t+1*?**
  - Follows a link uniformly at random

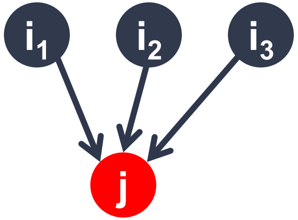
$$p(t+1) = M \cdot p(t)$$

- Suppose the random walk reaches a state

$$p(t+1) = M \cdot p(t) = p(t)$$

then $p(t)$ is **stationary distribution** of a random walk

- **Our original rank vector** r satisfies $r = M \cdot r$
  - **So, r is a stationary distribution for the random walk**



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

Page RanK: Google Formulations

Google Formulation

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

or equivalently

$$r = Mr$$

Google Formulation

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

or equivalently

$$r = Mr$$

Does this value converge ?

Google Formulation

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

or
equivalently

$$r = Mr$$

Does this value converge ?

Does it converge to the results that we want?

Google Formulation

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

or equivalently

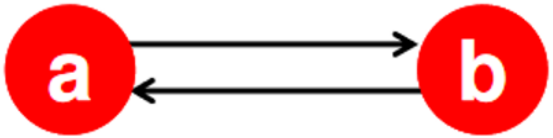
$$r = Mr$$

Does this value converge ?

Does it converge to the results that we want?

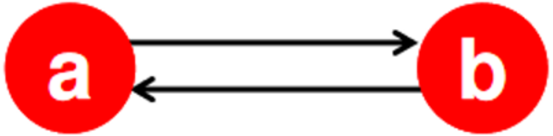
Are the results reasonable?

Does this converge?



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

Does this converge?



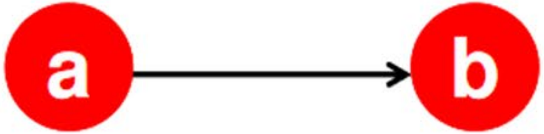
$$r_j^{(t+1)} = \sum_{i \to j} \frac{r_i^{(t)}}{\mathrm{d}_i}$$

- **Example:**

$$\begin{array}{c} r_a \\ r_b \end{array} = \begin{array}{cccc} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{array}$$

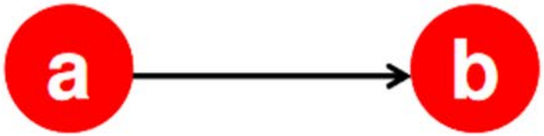
Iteration 0, 1, 2, …

Does this converge to what we want?



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

Does this converge to what we want?



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

- **Example:**

$$\begin{array}{ccccc} r_a & = & 1 & 0 & 0 & 0 \\ r_b & & 0 & 1 & 0 & 0 \end{array}$$

Iteration 0, 1, 2, …

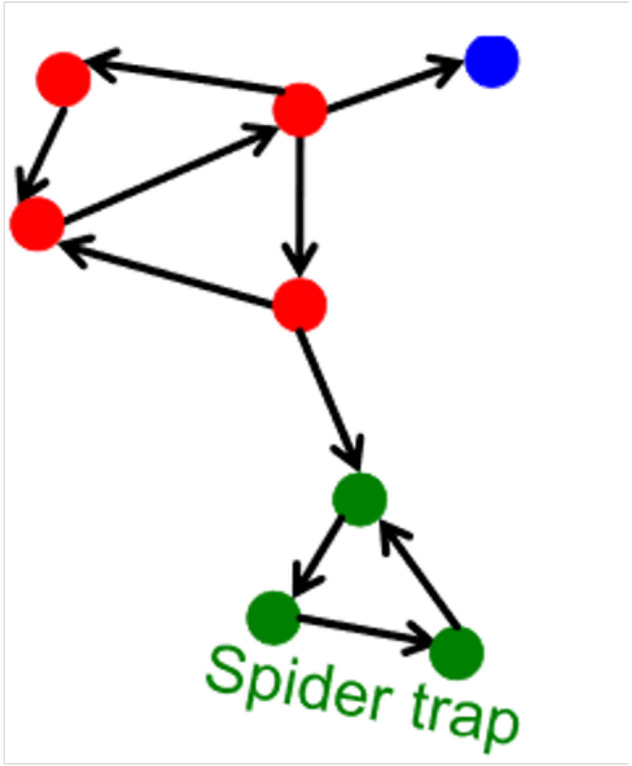
Page Rank: Two Problems

1. Dead ends:

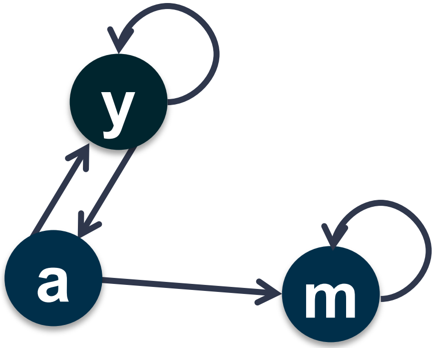
- Some pages are **dead ends** (have no out-link)
- Such pages cause important information to leak

2. Spider traps

- All out-links are within the group
- Random walk gets stuck in a trap
- And eventually spider traps absorbs all importance



Spider Traps



m is a spider trap

	y	a	m
y	½	½	0
a	½	0	0
m	0	½	1

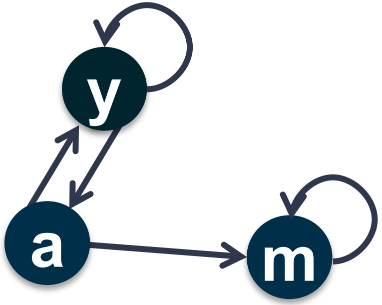
$$r_y = r_y/2 + r_a/2$$

$$r_a = r_y/2$$

$$r_m = r_a/2 + r_m$$

Spider Traps

- **Power Iteration:**
  - Set $r_j = 1$
  - $r_j = \sum_{i \to j} \frac{r_i}{d_i}$
    - And iterate



m is a spider trap

	y	a	m
y	½	½	0
a	½	0	0
m	0	½	1

$$\mathbf{r_y} = \mathbf{r_y}/2 + \mathbf{r_a}/2$$
$$\mathbf{r_a} = \mathbf{r_y}/2$$
$$\mathbf{r_m} = \mathbf{r_a}/2 + \mathbf{r_m}$$

- **Example:**

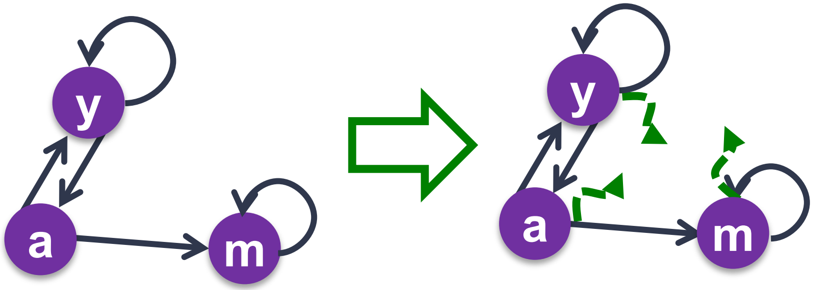
$$\begin{bmatrix} r_y \\ r_a \\ r_m \end{bmatrix} = \begin{matrix} 1/3 \\ 1/3 \\ 1/3 \end{matrix} \quad \begin{matrix} 2/6 \\ 1/6 \\ 3/6 \end{matrix} \quad \begin{matrix} 3/12 \\ 2/12 \\ 7/12 \end{matrix} \quad \begin{matrix} 5/24 \\ 3/24 \\ 16/24 \end{matrix} \quad \ldots \quad \begin{matrix} 0 \\ 0 \\ 1 \end{matrix}$$

Iteration 0, 1, 2, …

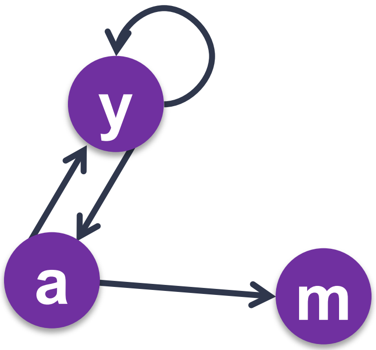
All the PageRank score gets “trapped” in node m.

Solution: Teleports!

- **The Google solution for spider traps: At each time step, the random surfer has two options**
  - With prob. β, follow a link at random
  - With prob. **1-**β, jump to some random page
  - Common values for β are in the range 0.8 to 0.9
  - *This will help the surfer to teleport out of spider trap within a few steps*



Dead Ends



	y	a	m
y	½	½	0
a	½	0	0
m	0	½	0

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

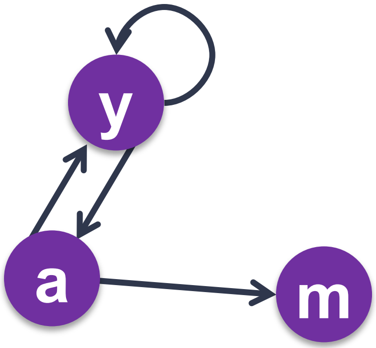
Dead Ends

Power Iteration:

Set $r_j = 1$

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

And iterate



	y	a	m
y	½	½	0
a	½	0	0
m	0	½	0

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

$$\begin{bmatrix} r_y \\ r_a \\ r_m \end{bmatrix} = \begin{matrix} 1/3 & 2/6 & 3/12 & 5/24 & & 0 \\ 1/3 & 1/6 & 2/12 & 3/24 & \ldots & 0 \\ 1/3 & 1/6 & 1/12 & 2/24 & & 0 \end{matrix}$$

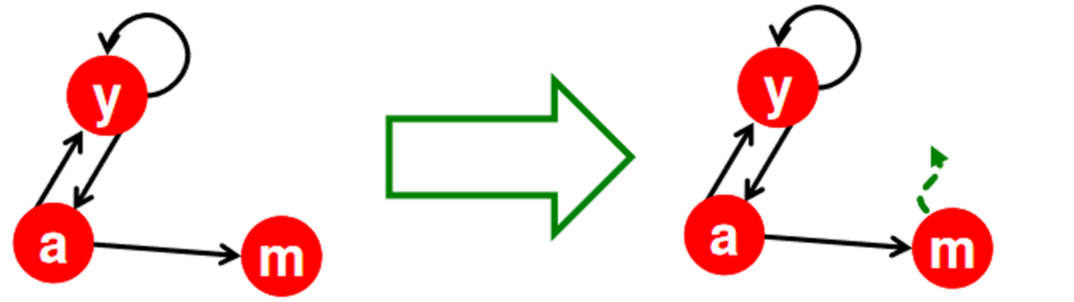
Iteration 0, 1, 2, …

Here the PageRank “leaks” out since the matrix is not stochastic.

Solution: Teleports

Teleport with probability 1.0 at dead ends

Adjust matric accordingly



	y	a	m
y	½	½	0
a	½	0	0
m	0	½	0

	y	a	m
y	½	½	⅓
a	½	0	⅓
m	0	½	⅓

Why Teleports Solve the Problem?

Why are dead-ends and spider traps a problem
and why do teleports solve the problem?

- **Spider-traps** are not a problem, but with traps PageRank scores are **not** what we want
    - **Solution:** Never get stuck in a spider trap by teleporting out of it in a finite number of steps
- **Dead-ends** are a problem
    - The matrix is not column stochastic so our initial assumptions are not met
    - **Solution:** Make matrix column stochastic by always teleporting when there is nowhere else to go

Google Solution: Random Teleports

- **Google's solution that does it all:**
  At each step, random surfer has two options:
  - With probability β, follow a link at random
  - With probability $1\text{-}\beta$, jump to some random page

- **PageRank equation** [Brin-Page, 98]

d_i … out-degree of node i

$$r_j = \sum_{i \to j} \beta \frac{r_i}{d_i} + (1-\beta)\frac{1}{N}$$

This formulation assumes that M has no dead ends. We can either preprocess matrix M to remove all dead ends or explicitly follow random teleport links with probability 1.0 from dead-ends.

The Google Matrix

- **PageRank equation** [Brin-Page, ‘98]

$$r_j = \sum_{i \to j} \beta \frac{r_i}{d_i} + (1-\beta)\frac{1}{N}$$

- **The Google Matrix *A*:**

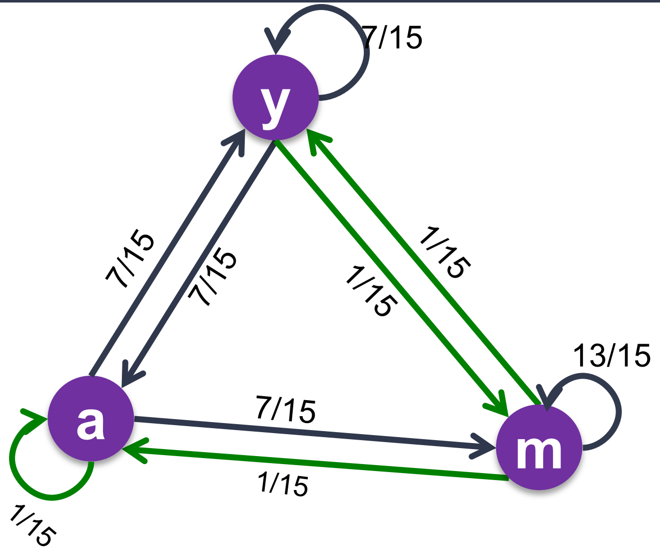
$$A = \beta\, M + (1-\beta)\left[\frac{1}{N}\right]_{N \times N}$$

$[1/N]_{NxN}$…N by N matrix where all entries are 1/N

- **We have a recursive problem:** $\boldsymbol{r} = \boldsymbol{A} \cdot \boldsymbol{r}$

- **And the Power method still works!**
- **What is β?**
  - In practice $\beta = 0.8, 0.9$ (make 5 steps on avg., jump)

Random Teleports ($\beta = 0.8$)



$$0.8 \overset{\mathbf{M}}{\begin{bmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 0 & 0 \\ 0 & 1/2 & 1 \end{bmatrix}} + 0.2 \overset{[1/N]_{N\times N}}{\begin{bmatrix} 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \end{bmatrix}}$$

$$\begin{array}{c} y \\ a \\ m \end{array} \underset{\mathbf{A}}{\begin{bmatrix} 7/15 & 7/15 & 1/15 \\ 7/15 & 1/15 & 1/15 \\ 1/15 & 7/15 & 13/15 \end{bmatrix}}$$

$$\begin{array}{c} y \\ a \\ m \end{array} = \begin{array}{c} 1/3 \\ 1/3 \\ 1/3 \end{array} \quad \begin{array}{c} 0.33 \\ 0.20 \\ 0.46 \end{array} \quad \begin{array}{c} 0.24 \\ 0.20 \\ 0.52 \end{array} \quad \begin{array}{c} 0.26 \\ 0.18 \\ 0.56 \end{array} \quad \ldots \quad \begin{array}{c} 7/33 \\ 5/33 \\ 21/33 \end{array}$$

References

[1] http://www.mmds.org

[2] Chapter 5 Lin and Dyer