Link Analysis

Shantanu Jain

PageRank

Using inbound links as “votes”

Credit: Yijun Zhao, Yi Wang,
Tan et al., Leskovec et al.

Web search before PageRank

- **Human-curated**
  (e.g. Yahoo, Looksmart)
  - Hand-written descriptions
  - Wait time for inclusion
- **Text-search**
  (e.g. WebCrawler, Lycos)
  - Prone to term-spam

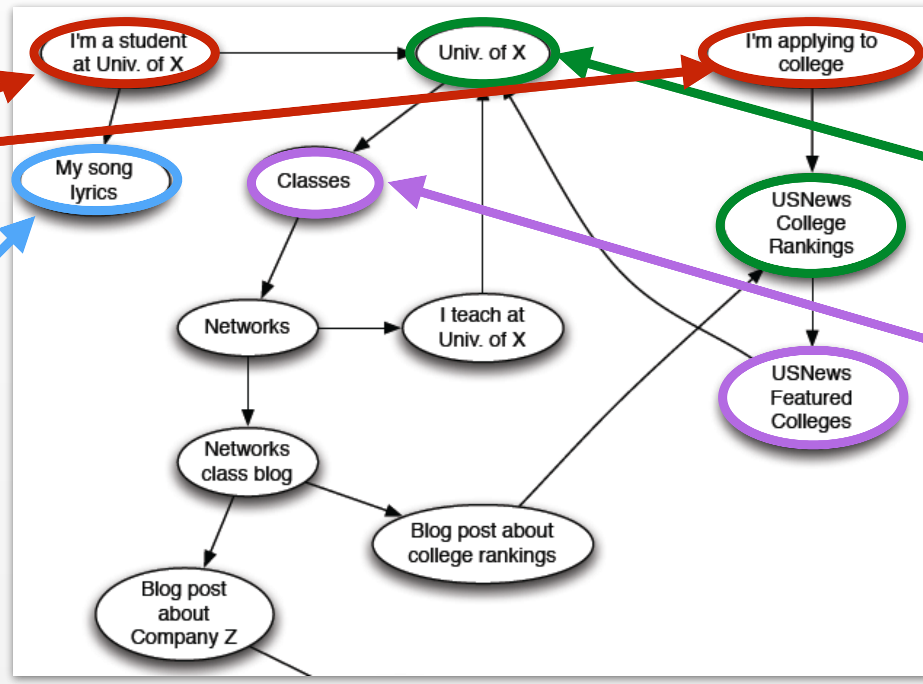
(adapted from:: Mining of Massive Datasets, http://www.mmds.org)

PageRank: Links as Votes

Not all pages are equally important

Few/no inbound links

Links from unimportant pages



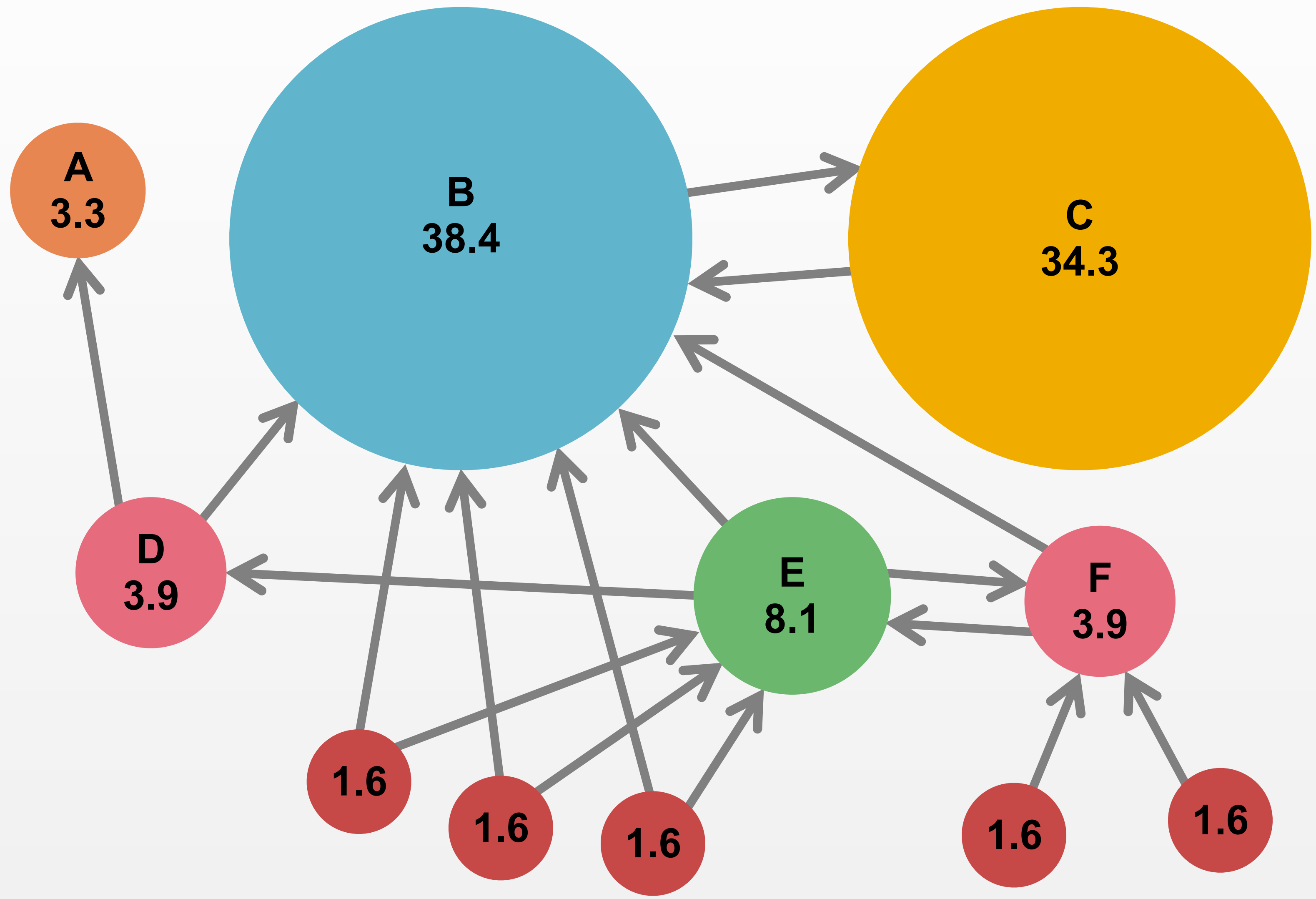
Many inbound links

Links from important pages

- Pages with **more inbound links** are more **important**
- Inbound **links from important pages** carry **more weight**

(adapted from:: Mining of Massive Datasets, http://www.mmds.org)

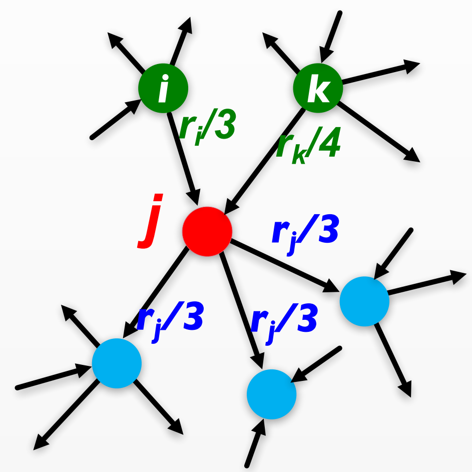
Example: PageRank Scores



(adapted from:: Mining of Massive Datasets, http://www.mmds.org)

PageRank: Recursive Formulation

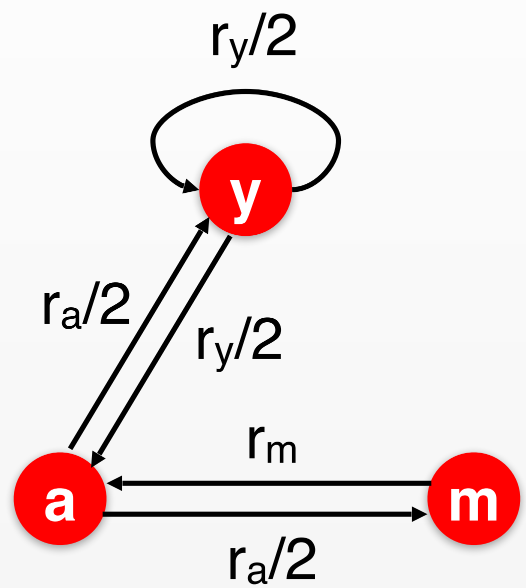
$$r_j = r_i/3 + r_k/4$$



- A link's vote is proportional to the importance of its source page
- If page j with importance r_j has n out-links, each link gets r_j / n votes
- Page j's own importance is the sum of the votes on its in-links

(adapted from:: Mining of Massive Datasets, http://www.mmds.org)

PageRank: The "Flow" Model



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

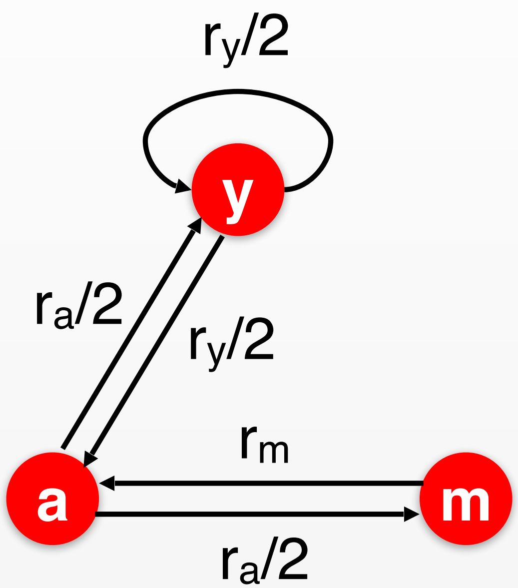
"Flow" equations:

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

- 3 equations, 3 unknowns
- However, the equations are under constrained. Has infinite solutions.
- Impose additional constraint: $r_y + r_a + r_m = 1$ (the total sum of importances is 1)
- Solution: $r_y = 2/5,\ r_a = 2/5,\ r_m = 1/5$

(adapted from:: Mining of Massive Datasets, http://www.mmds.org)

PageRank: The "Flow" Model



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

"Flow" equations:

$r_y = r_y/2 + r_a/2$

$r_a = r_y/2 + r_m$

$r_m = r_a/2$

r* = *M* · *r

$$\begin{bmatrix} r_y \\ r_a \\ r_m \end{bmatrix} = \begin{array}{c} \begin{array}{ccc} y & a & m \end{array} \\ \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{2} & 0 & 1 \\ 0 & \frac{1}{2} & 0 \end{bmatrix} \end{array} \begin{bmatrix} r_y \\ r_a \\ r_m \end{bmatrix}$$

Matrix *M* is stochastic (i.e. columns sum to one)

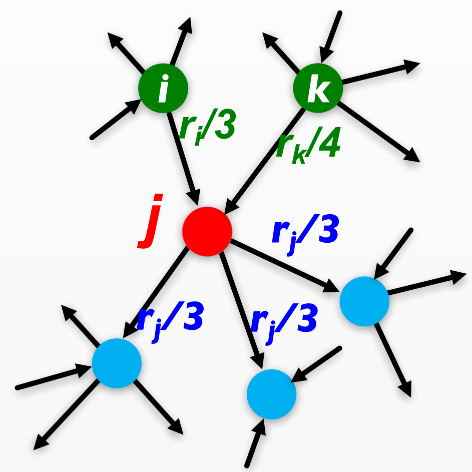
(adapted from:: Mining of Massive Datasets, http://www.mmds.org)

PageRank: Eigenvector Problem

- PageRank: Solve for eigenvector $r = M\, r$ with eigenvalue $\lambda = 1$.
- Normalize the eigenvector to sum to 1.
- Eigenvector with $\lambda = 1$ is guaranteed to exist since M is a stochastic matrix (i.e. if $a = M\, b$ then $\Sigma\, a_i = \Sigma\, b_i$)
- *Problem*: There are billions of pages on the internet. How do we solve for eigenvector with order 10^{10} elements?

Equivalent Formulation: Random Surfer

$r_j = r_i/3 + r_k/4$



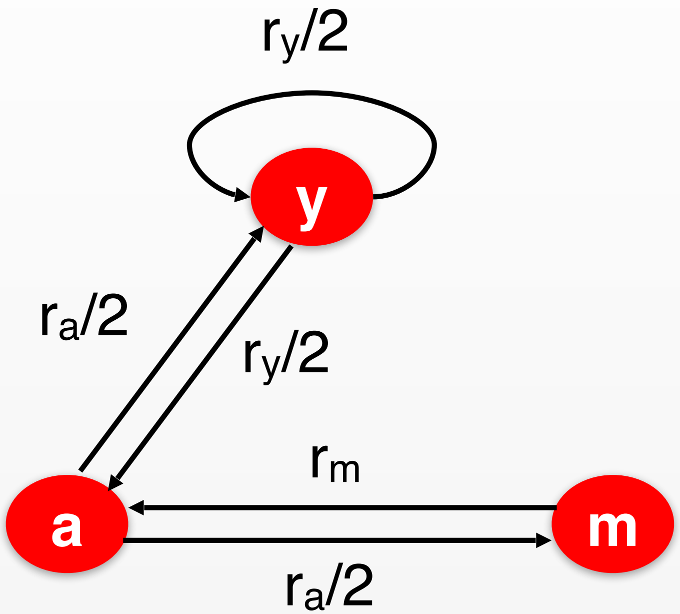
- At time *t* a surfer is on some page *i*
- At time *t+1* the surfer follows a link to a new page at random
- Define rank r_i as fraction of time spent on page *i*

(adapted from:: Mining of Massive Datasets, http://www.mmds.org)

PageRank: Power Iteration

Model for random Surfer:

- At time $t = 0$ pick a page at random
- At each subsequent time t follow an outgoing link at random



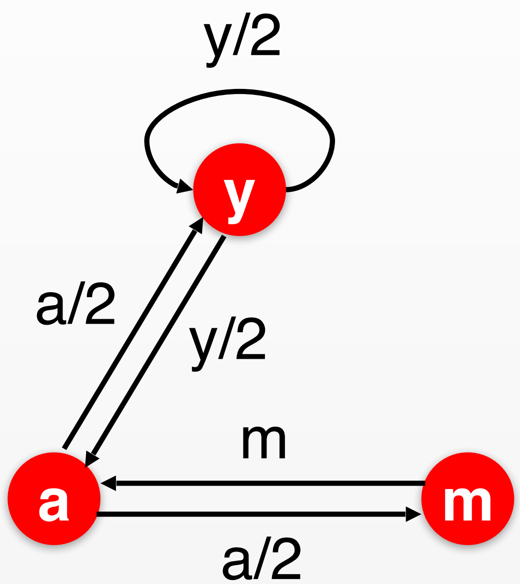
Probabilistic interpretation:

$$p(z_0 = i) = 1/N$$
$$p(z_t = i \,|\, z_{t-1} = j) = M_{ij}$$
$$p(z_t = i) = \textstyle\sum_j p(z_t = i, z_{t-1} = j)$$
$$= \textstyle\sum_j M_{ij} p(z_{t-1} = j)$$

z_i is the random variable giving the index of the webpage the random surfer is at time step i.

$$M = \begin{array}{c} \begin{array}{ccc} \text{y} & \text{a} & \text{m} \end{array} \\ \left[\begin{array}{ccc} \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{2} & 0 & 1 \\ 0 & \frac{1}{2} & 0 \end{array}\right] \end{array}$$

PageRank: Power Iteration



$$\boldsymbol{p}^t = M\boldsymbol{p}^{t-1} = M^t\boldsymbol{p}^0$$

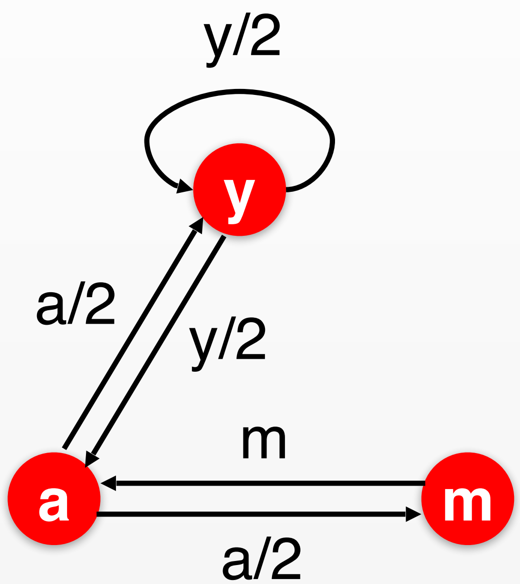
$$\boldsymbol{p}^0 = \begin{bmatrix} 1/3 \\ 1/3 \\ 1/3 \end{bmatrix} \quad M = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{2} & 0 & 1 \\ 0 & \frac{1}{2} & 0 \end{bmatrix}$$

$$\underset{p^0}{\begin{bmatrix} 1/3 \\ 1/3 \\ 1/3 \end{bmatrix}} \quad \underset{p^1 = Mp^0}{\begin{bmatrix} 2/6 \\ 3/6 \\ 1/6 \end{bmatrix}} \quad \underset{p^2 = MP^1}{\begin{bmatrix} 5/12 \\ 4/12 \\ 3/12 \end{bmatrix}} \quad \underset{p^3 = MP^2}{\begin{bmatrix} 9/24 \\ 11/24 \\ 4/24 \end{bmatrix}} \quad \underset{p^4 = MP^3}{\begin{bmatrix} 20/48 \\ 17/48 \\ 11/48 \end{bmatrix}} \simeq \begin{bmatrix} 2/5 \\ 2/5 \\ 1/5 \end{bmatrix}$$

p^t converges to r. Iterate until $|p^t - p^{t-1}| < \varepsilon$

(adapted from:: Mining of Massive Datasets, http://www.mmds.org)

PageRank: Convergence



$$\boldsymbol{p}^t = M\boldsymbol{p}^{t-1} = M^t\boldsymbol{p}^0$$

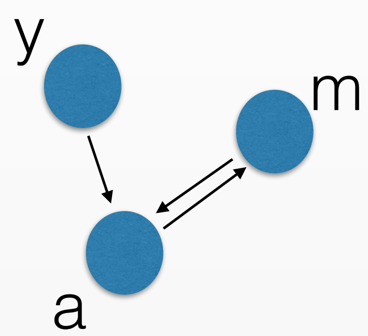
p^t converges to r. Iterate until $|p^t - p^{t-1}| < \varepsilon$

Is convergence always guaranteed?

Is the value at convergence meaningful?

(adapted from:: Mining of Massive Datasets, http://www.mmds.org)

Case of non-convergence



periodicity in the network

$$\boldsymbol{p}^t = M\boldsymbol{p}^{t-1} = M^t\boldsymbol{p}^0$$

$$\boldsymbol{p}^0 = \begin{bmatrix} 1/3 \\ 1/3 \\ 1/3 \end{bmatrix} \qquad M = \begin{array}{c} \begin{array}{ccc} \text{y} & \text{a} & \text{m} \end{array} \\ \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} \end{array}$$

a

$$\underset{p^0}{\begin{bmatrix} \frac{1}{3} \\ \frac{1}{3} \\ \frac{1}{3} \end{bmatrix}} \quad \underset{p^1}{\begin{bmatrix} 0 \\ \frac{2}{3} \\ \frac{1}{3} \end{bmatrix}} \quad \underset{p^2}{\begin{bmatrix} 0 \\ \frac{1}{3} \\ \frac{2}{3} \end{bmatrix}} \quad \underset{p^3}{\begin{bmatrix} 0 \\ \frac{2}{3} \\ \frac{1}{3} \end{bmatrix}} \quad \underset{p^4}{\begin{bmatrix} 0 \\ \frac{1}{3} \\ \frac{2}{3} \end{bmatrix}}$$

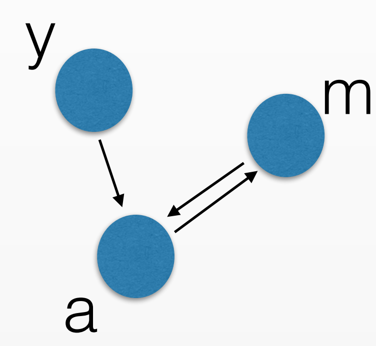
What if you started from?

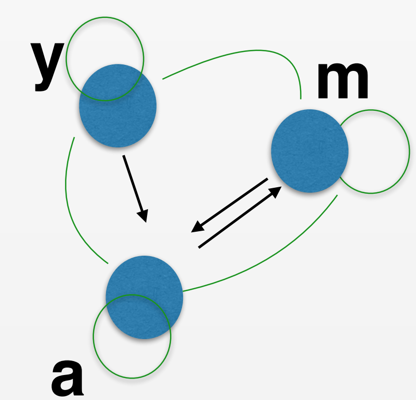
$$p^0 = \begin{bmatrix} 0 \\ \frac{1}{2} \\ \frac{1}{2} \end{bmatrix}$$

Solution: Random Teleports

Model for teleporting random surfer:

- At time $t = 0$ pick a page at random
- At each subsequent time t
  - With probability β follow an outgoing link at random
  - With probability 1-β teleport to a new initial location at random



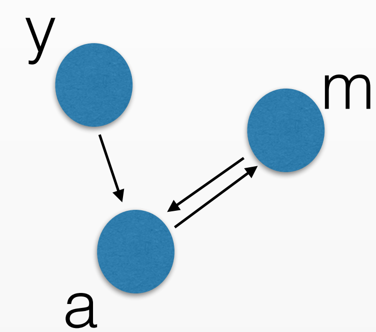


$$p(z_{t+1} = i \mid z_t = j) = \beta \frac{1}{d_j} + (1 - \beta) \frac{1}{N}$$

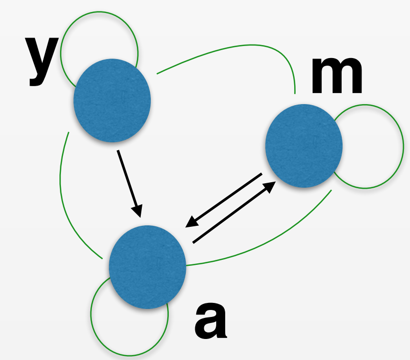
(adapted from:: Mining of Massive Datasets, http://www.mmds.org)

Solution: Random Teleports

$$p(z_{t+1} = i \,|\, z_t = j) = \beta \frac{1}{d_j} + (1-\beta)\frac{1}{N}$$



$$\tilde{M} = 0.9 \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} + 0.1 \begin{bmatrix} 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \end{bmatrix} = \begin{array}{c} \begin{array}{ccc} \text{y} & \text{a} & \text{m} \end{array} \\ \begin{bmatrix} .033 & .033 & .033 \\ .933 & .033 & .933 \\ .033 & .933 & .033 \end{bmatrix} \end{array}$$

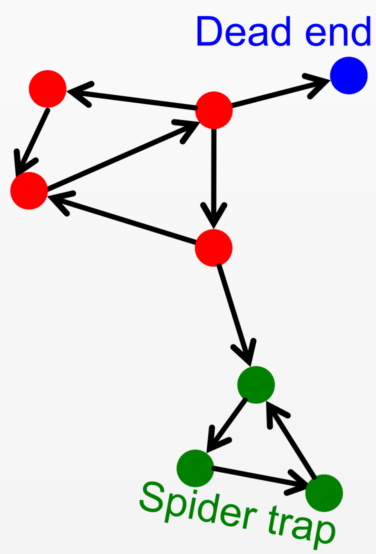


$$p^{t+1} = \tilde{M} p^t$$

$$\underset{p^0}{\begin{bmatrix} 1/3 \\ 1/3 \\ 1/3 \end{bmatrix}} \dots \underset{p^{100}}{\begin{bmatrix} .033 \\ .491 \\ .475 \end{bmatrix}} \quad \underset{p^{101}}{\begin{bmatrix} .033 \\ .491 \\ .475 \end{bmatrix}} \quad \underset{p^{102}}{\begin{bmatrix} .033 \\ .491 \\ .475 \end{bmatrix}}$$

(adapted from:: Mining of Massive Datasets, http://www.mmds.org)

PageRank: Problems



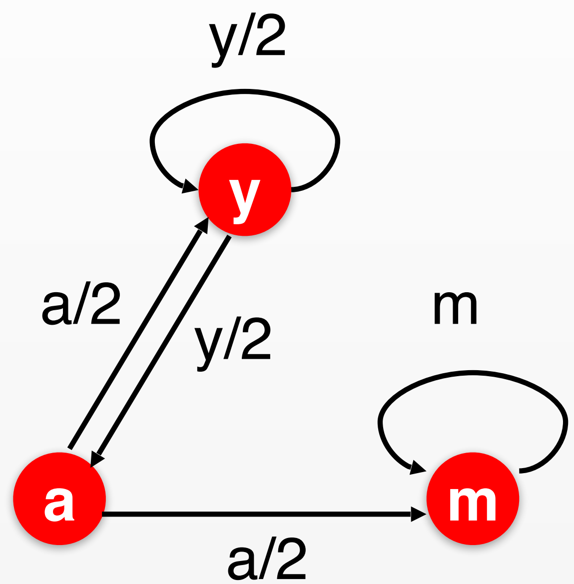
Not irreducible

1. *Dead Ends*
   - Nodes with no outgoing links.
   - Where do surfers go next?

2. *Spider Traps*
   - Subgraph with no outgoing links to wider graph
   - Surfers are "trapped" with no way out.

(adapted from:: Mining of Massive Datasets, http://www.mmds.org)

Power Iteration: Spider Traps



$$\boldsymbol{p}^t = M\boldsymbol{p}^{t-1} = M^t\boldsymbol{p}^0$$

$$\boldsymbol{p}^0 = \begin{bmatrix} 1/3 \\ 1/3 \\ 1/3 \end{bmatrix} \quad M = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{2} & 0 & 0 \\ 0 & \frac{1}{2} & 1 \end{bmatrix}$$

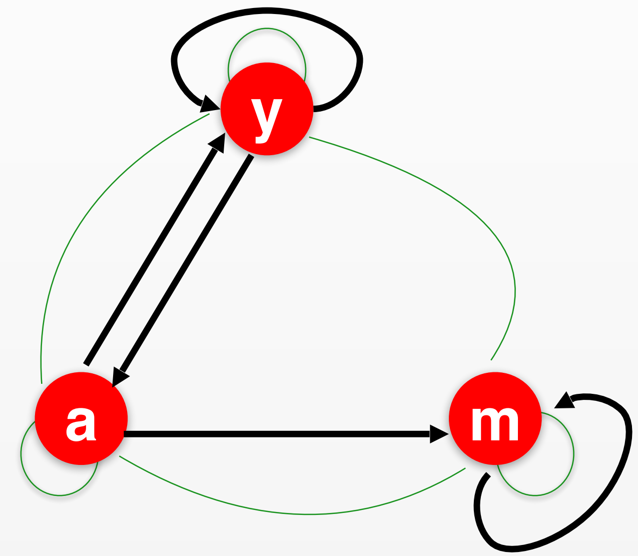
$$\underset{p^0}{\begin{bmatrix} 1/3 \\ 1/3 \\ 1/3 \end{bmatrix}} \underset{p^1}{\begin{bmatrix} 2/6 \\ 1/6 \\ 3/6 \end{bmatrix}} \underset{p^2}{\begin{bmatrix} 3/12 \\ 2/12 \\ 7/12 \end{bmatrix}} \underset{p^3}{\begin{bmatrix} 5/24 \\ 3/24 \\ 16/24 \end{bmatrix}} \ldots \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

Probability accumulates in traps (surfers get stuck)

(adapted from:: Mining of Massive Datasets, http://www.mmds.org)

Power Iteration: Teleports

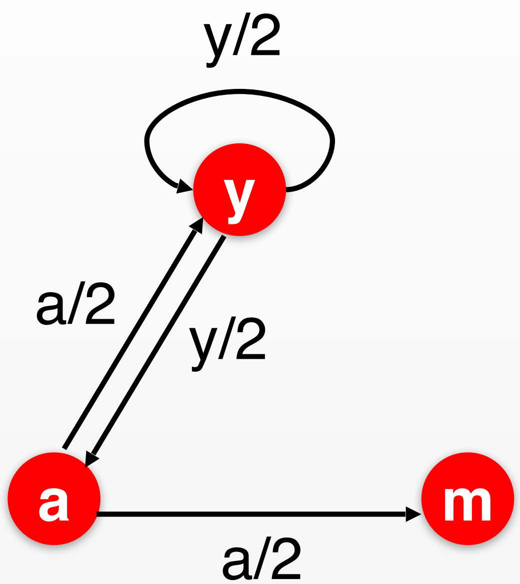
$$\tilde{M} = 4/5 \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{2} & 0 & 0 \\ 0 & \frac{1}{2} & 1 \end{bmatrix} + 1/5 \begin{bmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix} = \begin{bmatrix} \frac{7}{15} & \frac{7}{15} & \frac{1}{15} \\ \frac{7}{15} & \frac{1}{15} & \frac{1}{15} \\ \frac{1}{15} & \frac{7}{15} & \frac{1}{15} \end{bmatrix}$$



$$\underset{p^0}{\begin{bmatrix} 1/3 \\ 1/3 \\ 1/3 \end{bmatrix}} \quad \underset{p^1}{\begin{bmatrix} 0.33 \\ 0.20 \\ 0.46 \end{bmatrix}} \quad \underset{p^2}{\begin{bmatrix} 0.24 \\ 0.20 \\ 0.56 \end{bmatrix}} \quad \dots \quad \begin{bmatrix} 7/33 \\ 5/33 \\ 21/33 \end{bmatrix}$$

(adapted from:: Mining of Massive Datasets, http://www.mmds.org)

Power Iteration: Dead Ends



$$\boldsymbol{p}^t = M\boldsymbol{p}^{t-1} = M^t\boldsymbol{p}^0$$

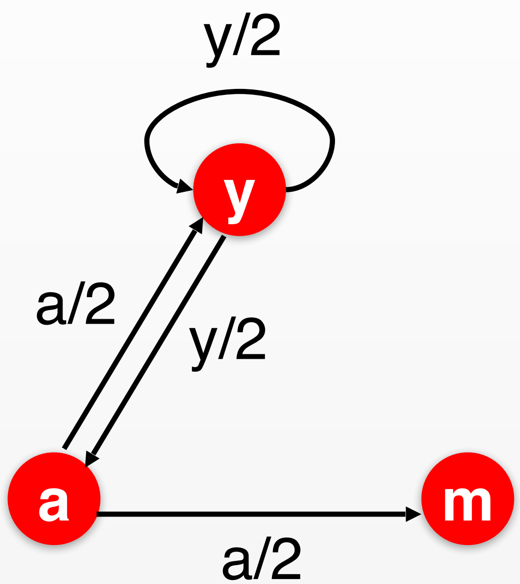
$$\boldsymbol{p}^0 = \begin{bmatrix} 1/3 \\ 1/3 \\ 1/3 \end{bmatrix} \quad M = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{2} & 0 & 0 \\ 0 & \frac{1}{2} & 0 \end{bmatrix}$$

$$\boldsymbol{p}^t = \underset{p^0}{\begin{bmatrix} 1/3 \\ 1/3 \\ 1/3 \end{bmatrix}} \underset{p^1}{\begin{bmatrix} 2/6 \\ 1/6 \\ 1/6 \end{bmatrix}} \underset{p^2}{\begin{bmatrix} 3/12 \\ 1/12 \\ 1/12 \end{bmatrix}} \ldots \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

Probability not conserved

(adapted from:: Mining of Massive Datasets, http://www.mmds.org)

Power Iteration: Dead Ends



$$\boldsymbol{p}^t = M\boldsymbol{p}^{t-1} = M^t\boldsymbol{p}^0$$

$$\boldsymbol{p}^0 = \begin{bmatrix} 1/3 \\ 1/3 \\ 1/3 \end{bmatrix} \quad M = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & \frac{1}{3} \\ \frac{1}{2} & 0 & \frac{1}{3} \\ 0 & \frac{1}{2} & \frac{1}{3} \end{bmatrix}$$

(teleport at dead ends)

$$\boldsymbol{p}^t = \begin{bmatrix} 1/3 \\ 1/3 \\ 1/3 \end{bmatrix} \begin{bmatrix} 8/18 \\ 5/18 \\ 5/18 \end{bmatrix} \begin{bmatrix} 49/108 \\ 34/108 \\ 35/108 \end{bmatrix} \ldots$$

Fixes "probability sink" issue

(adapted from:: Mining of Massive Datasets, http://www.mmds.org)

Power Iteration: Dead Ends

M: the original probability transition matrix

$\ddot{M}$: Constructed from M by replacing any column corresponding to a dead end (contains only 0's) by a column containing all entries equal to1/N.

$$\tilde{M} = \beta \ddot{M} + (1 - \beta)\frac{1}{N}\mathbf{1}\mathbf{1}^T$$

$$\mathbf{1} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$$

(adapted from:: Mining of Massive Datasets, http://www.mmds.org)

Intermezzo: Markov Chains

Markov Property

$$p(z_{t+1} \mid z_t) = p(z_{t+1} \mid z_t, z_{t-1}, \ldots z_0)$$

Irreducibility

$$\forall i, j \; \exists n_{ij} > 0 : \; p(z_{\eta_{ij}} = \mid z_0 = j) > 0$$

Ergodicity

$$\exists N : N \geq \max_{i,j} n_{ij}$$

Aperiodicity:

$$\exists i, \text{GCD}(T_i) = 1 \qquad T_i = \{t \geq 1 : p(z_t = i \mid z_0 = i) > 0)$$

There exists r with all entries strictly positive

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

Power iteration converges to r, starting with any probability vector p^0

$$\lim_{t \to \infty} \boldsymbol{p}^t = \boldsymbol{r}$$

where

$$\boldsymbol{p}^t = M\boldsymbol{p}^{t-1}$$