Analytics and Visualization of Big Data

Fadel M. Megahed

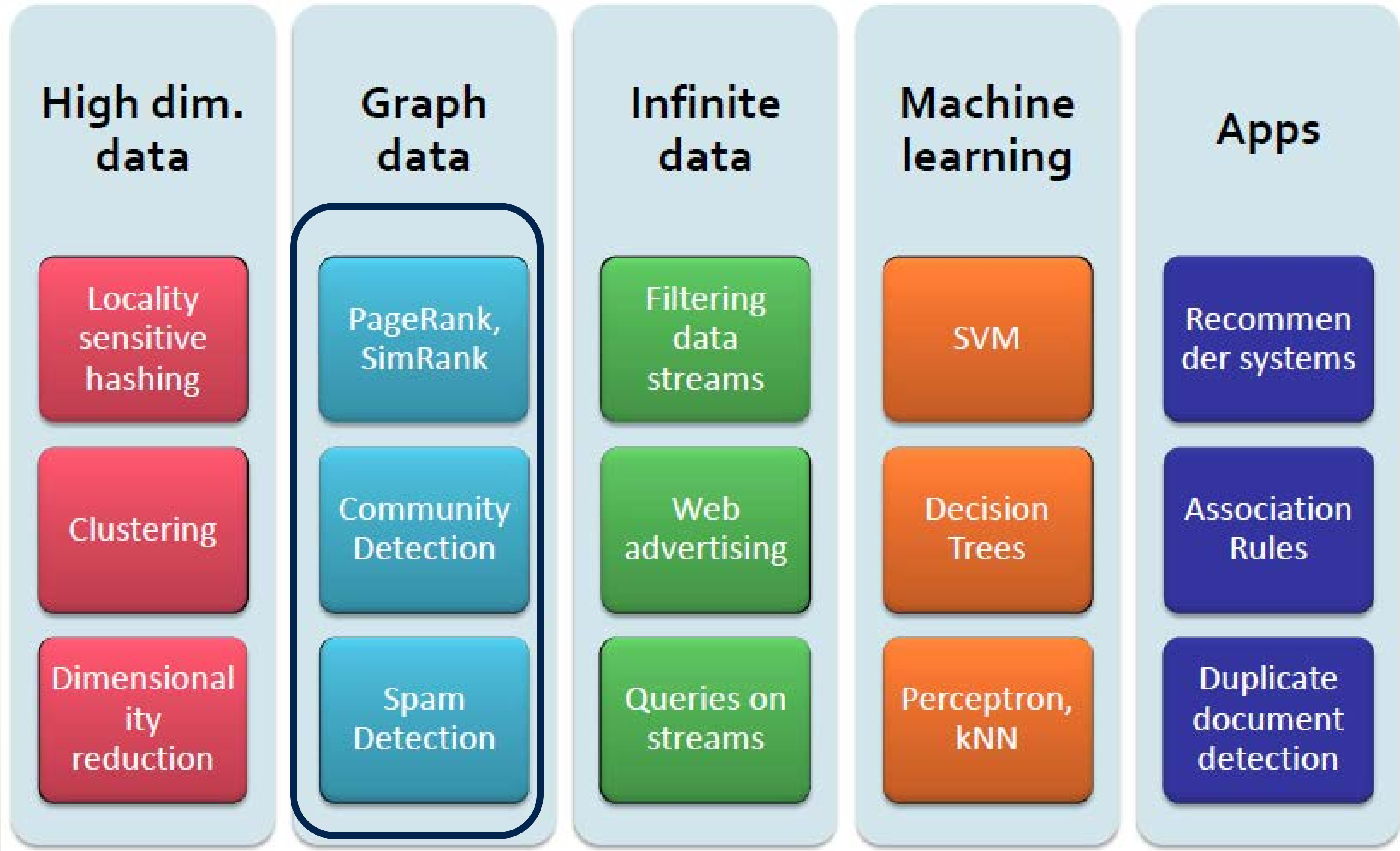
Lecture 20: Link Analysis (Cont.)

AUBURN UNIVERSITY
SAMUEL GINN
COLLEGE OF ENGINEERING

Department of Industrial and Systems Engineering

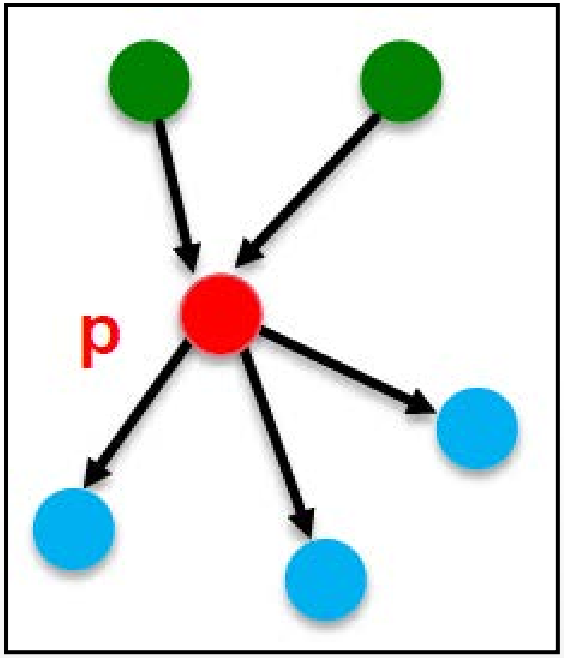
Spring 13

Refresher: Analytics Based on Data Type



Refresher: The Brilliant Idea that Made Google – PageRank

- **Idea: Links as votes**
  - Page is more important if it has more links
    - In-coming links? Out-going links?
- **Think of in-links as votes**:
  - www.auburn.edu
  - www.joe-schmoe.com
- **Are all in-links are equal?**
  - Links from important pages count more
  - Recursive question!



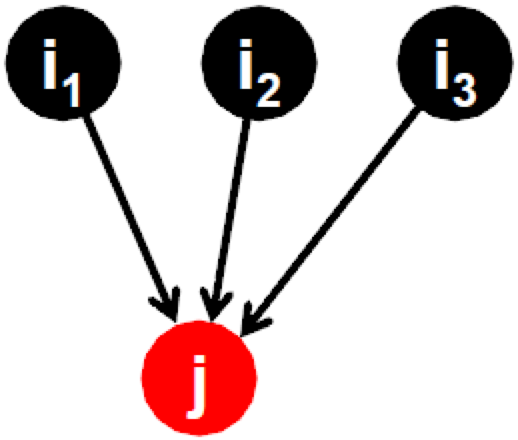
Refresher: What do we mean by recursive?

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

The Interpretation of Our Formulation

- **Imagine a random web surfer:**
  - At any time t, 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 indefinitely

- **Let:**
  - $\boldsymbol{p}(\boldsymbol{t})$ … vector whose ith coordinate is the prob. that the surfer is at page i at time t
  - So, $\boldsymbol{p}(\boldsymbol{t})$ is a probability distribution over pages



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

Source: Jure Leskovic, Stanford CS246, Lecture Notes, see http://cs246.stanford.edu

The Stationary Distribution

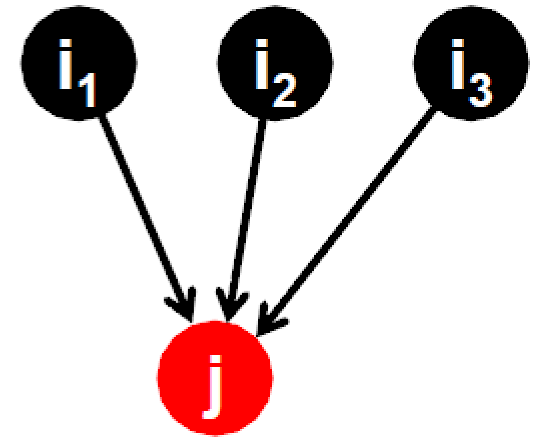
- **Where is the surfer at time *t+1*?**
  - Follows a link uniformly at random

    $$p\,(t+1) = M * p(t)$$

  - Suppose the random walk reaches a state

    $$p\,(t+1) = M * p(t) = p(t)$$

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



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

- **Our original rank vector *r*** satisfies that since $\boldsymbol{r} = \boldsymbol{M} * \boldsymbol{r}$
  - **So, *r* is a stationary distribution for the random walk**

Source: Jure Leskovic, Stanford CS246, Lecture Notes, see http://cs246.stanford.edu

Three Questions that Will be Addressed in Today's Class

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

or equivalently

$$r = Mr$$

1. Does this converge?
2. Does it converge to what we want?
3. Are results reasonable?

Does this converge?

Does this converge to what we want?

Exercise: Based on our discussion from last class, please answer these two questions

Problems with the Flow Model

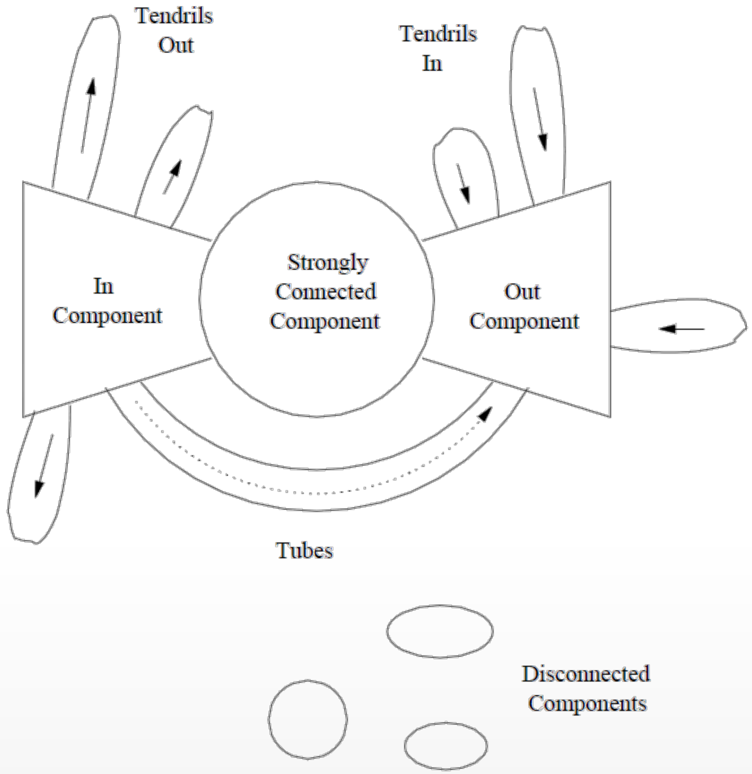


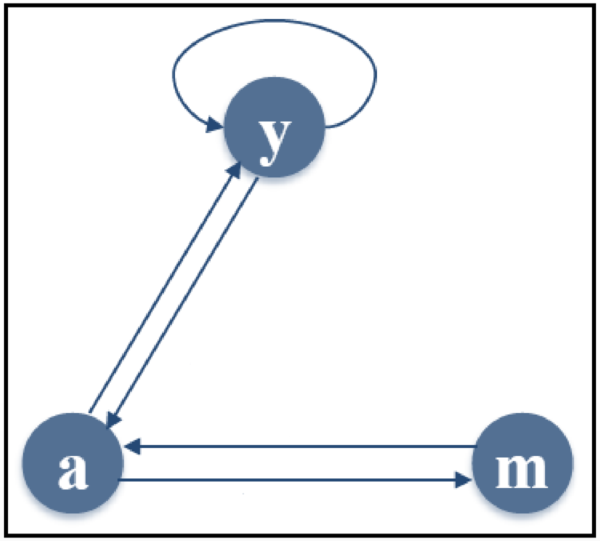
Figure 5.2: The "bowtie" picture of the Web

There exists two problems with the flow model:

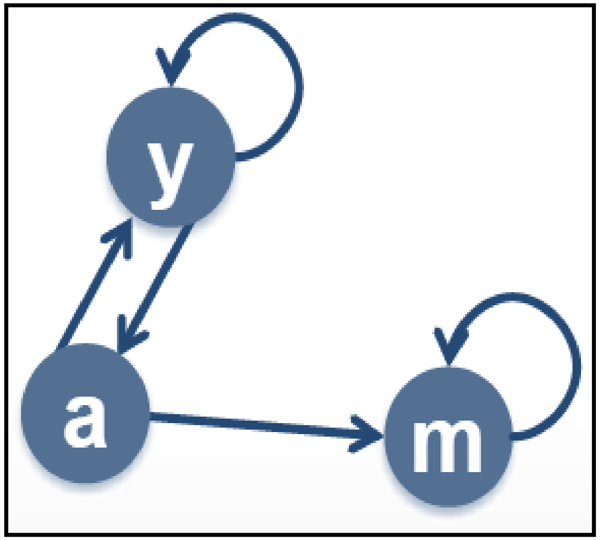
1. Some pages are "**dead ends**"
   - Such pages cause importance to "leak out"
2. **Spider Traps**
   - Eventually, they absorb all importance

The "Spider Trap" Problem

Example from Last Class



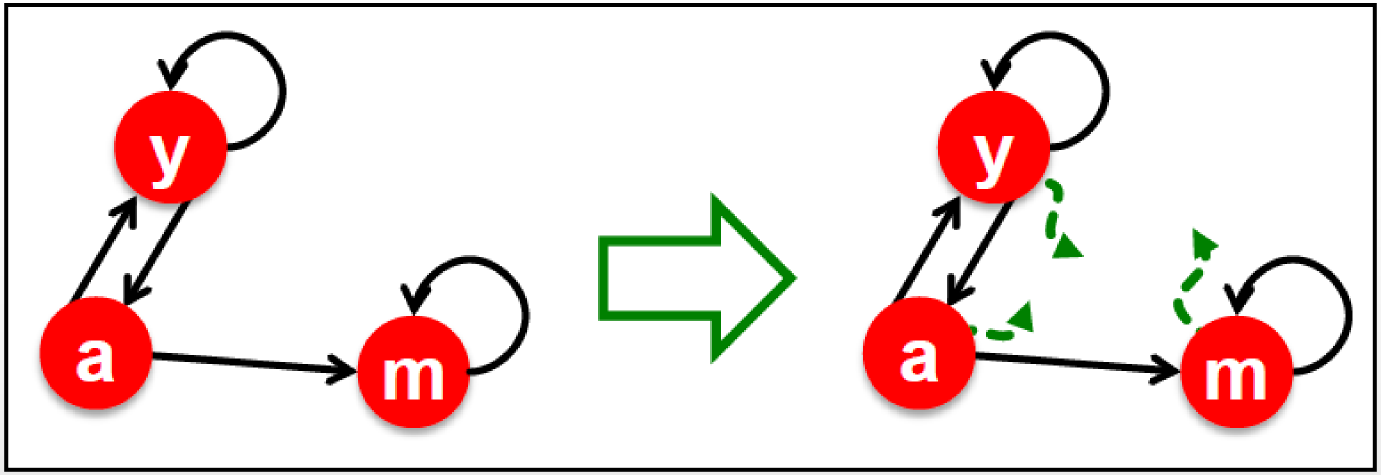
Modified Example



Let us work it out together to see the difference in Convergence ☺

Solution: Random Teleport

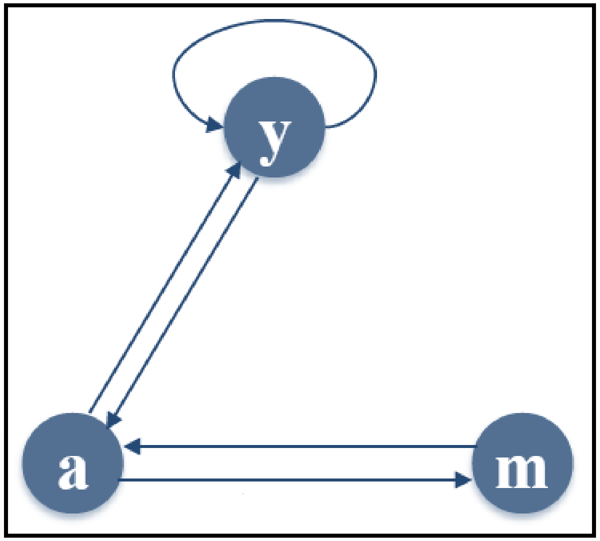
- **The Google solution for spider traps: At each time step, the random surfer has two options:**
  - With probability β, follow a link at random
  - With probability 1-β, jump to some page uniformly at random
  - Common values for β are in the range 0.8 to 0.9
- **Surfer will teleport out within a few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



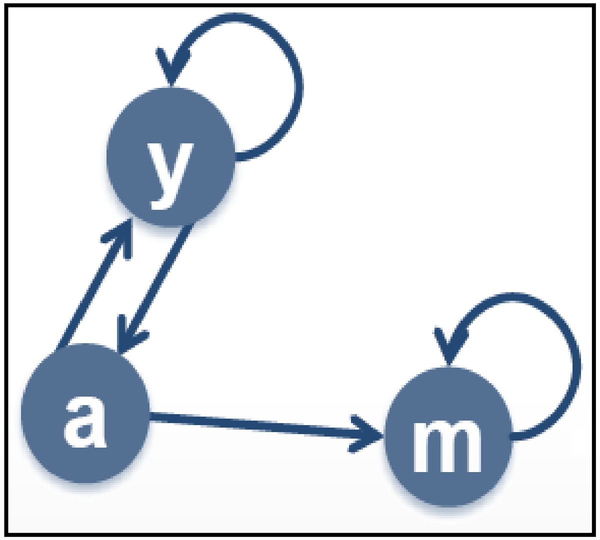
Source: Jure Leskovic, Stanford CS246, Lecture Notes, see http://cs246.stanford.edu

The "Dead-End" Problem

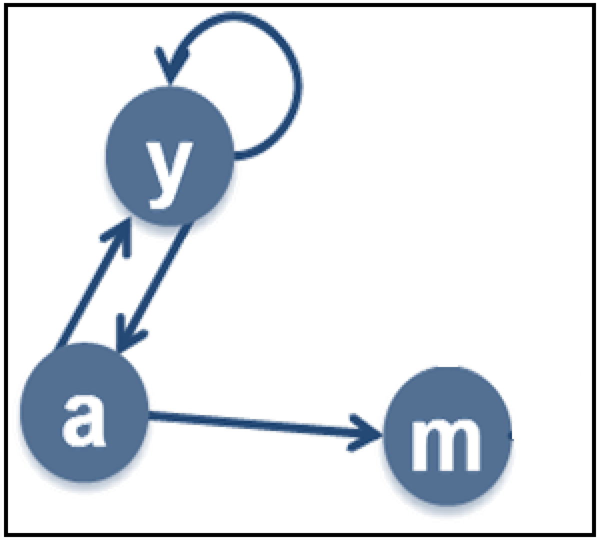
Standard Example



Spider-Web Example



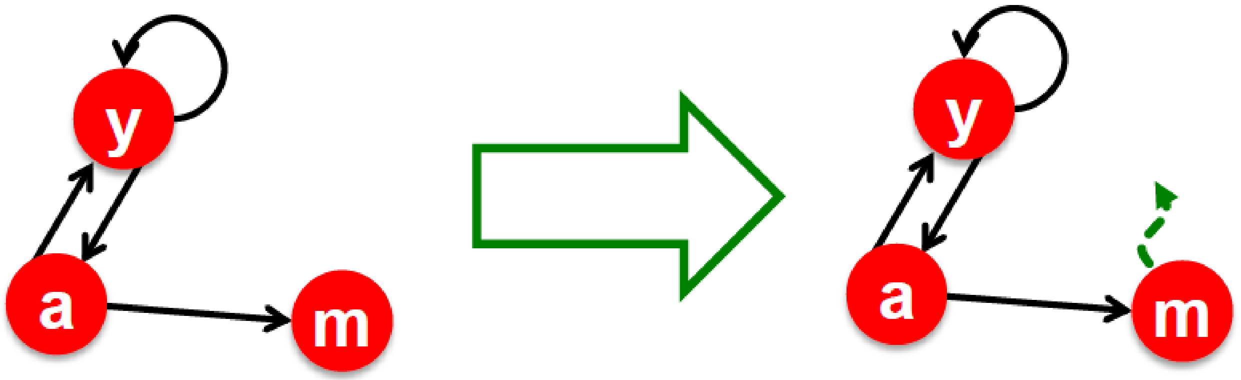
Dead-End



What is the impact of dead-end on the convergence of the **r** vector?

Solution: Teleport

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

Source: Jure Leskovic, Stanford CS246, Lecture Notes, see http://cs246.stanford.edu

Why Teleports Solve the Problem?

$$r^{(t+1)} = Mr^{(t)}$$

Markov Chains

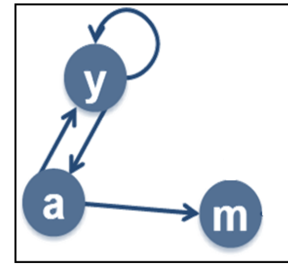
- Set of states X
- Transition matrix P where $P_{ij} = P(X_t=i \mid X_{t-1}=j)$
- π specifying the probability of being at each state x ∈ X
- Goal is to find π such that π = P π

Theory of Markov Chains

- For any start vector, the power method applied to a transition matrix P will converge to a unique positive stationary vector as long as P is **stochastic**, **irreducible** and **aperiodic**.

Making M Stochastic

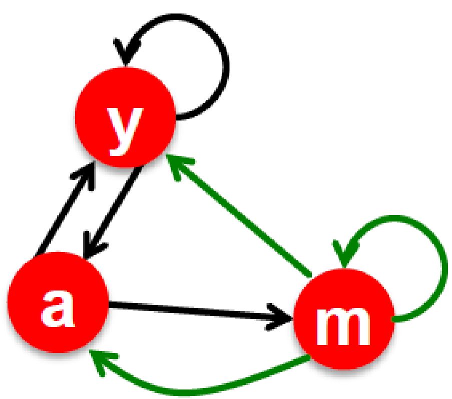
- **Stochastic:** Every column sums to 1



- **A possible solution:** Add green links

$$S = M + a^T \left(\frac{1}{n}\mathbf{1}\right)$$

- a_i…=1 if node *i* has out deg 0, =0 else
- **1**…vector of all 1s

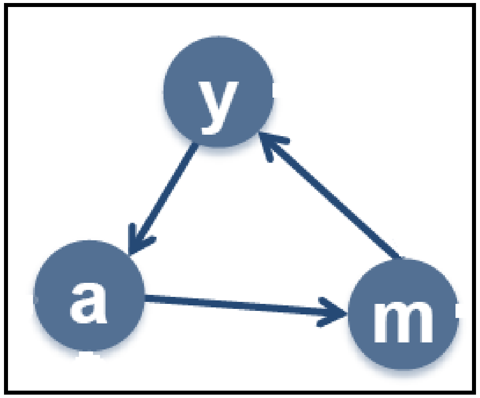


	y	a	m
y	½	½	1/3
a	½	0	1/3
m	0	½	1/3

$$r_y = r_y/2 + r_a/2 + r_m/3$$
$$r_a = r_y/2 + r_m/3$$
$$r_m = r_a/2 + r_m/3$$

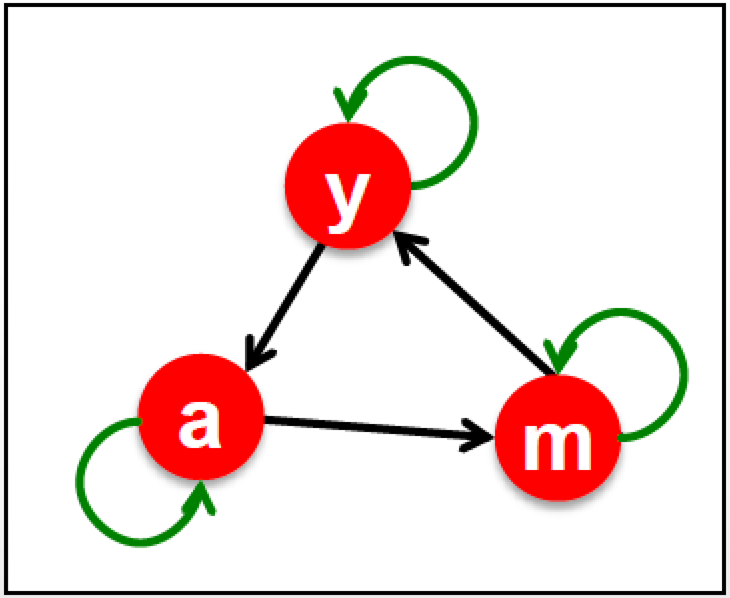
Source: Jure Leskovic, Stanford CS246, Lecture Notes, see http://cs246.stanford.edu

Make M Aperiodic

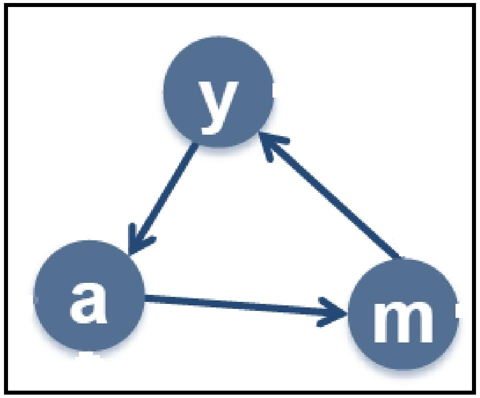


- A chain is **periodic** if there exists $k > 1$ such that the interval between two visits to some state s is always a multiple of k.

- **A possible solution:** Add **green** links

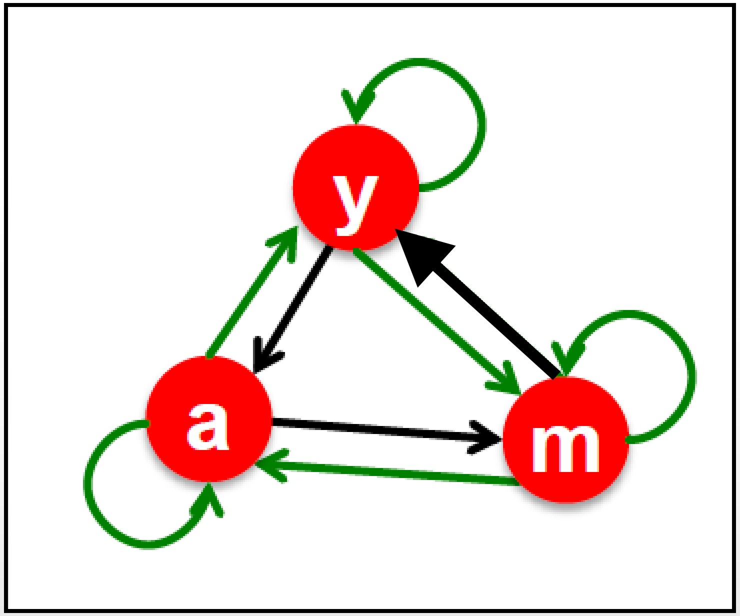


Make M Irreducible



- **Definition:** From any state, there is a non-zero probability of going from any one state to any another

- **A possible solution:** Add green links



Solution: Random Jumps

- **Google's solution that does it all:**
  - Makes ***M*** stochastic, aperiodic, irreducible
- **At each step, random surfer has two options:**
  - With probability *1*-β, follow a link at random
  - With probability β, jump to some random page
- **PageRank equation** [Brin-Page, 98]

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

Assuming we follow random teleport links with probability 1.0 from dead-ends

d_i ... out-degree of node i

Source: Jure Leskovic, Stanford CS246, Lecture Notes, see http://cs246.stanford.edu

In-depth Discussion (FYI): The Google Matrix

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

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

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

$$A = (1-\beta)S + \beta\frac{1}{n}\mathbf{1}\cdot\mathbf{1}^T$$

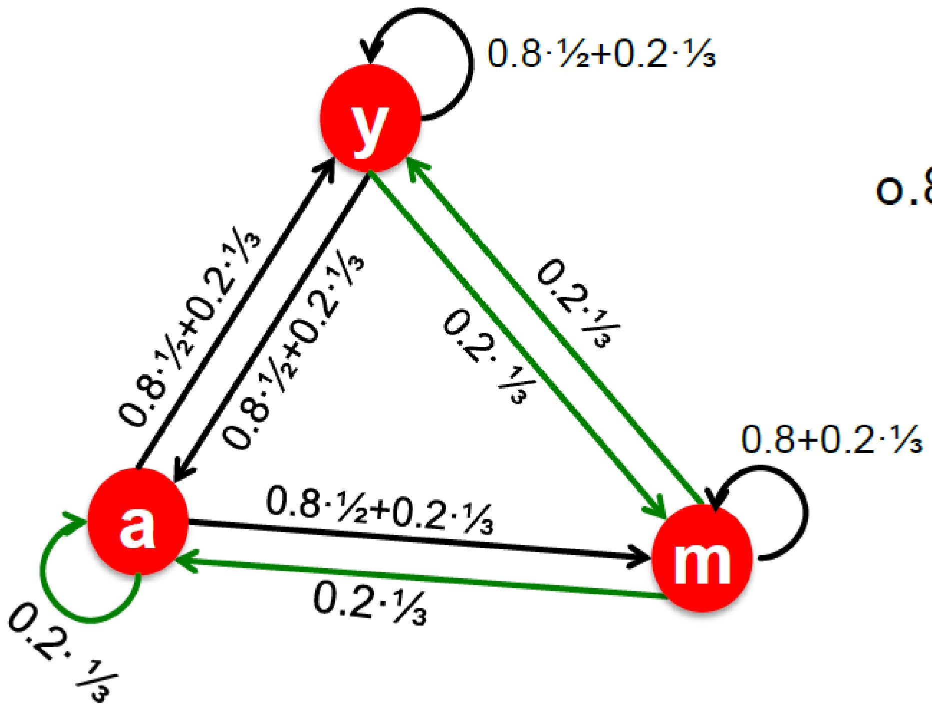
- ***G* is stochastic, aperiodic and irreducible, so**

$$r^{(t+1)} = A \cdot r^{(t)}$$

- **What is β ?**
  - In practice $\beta = 0.15$ (make *5* steps and jump)

Source: Jure Leskovic, Stanford CS246, Lecture Notes, see http://cs246.stanford.edu

In-depth Discussion (FYI): An Example



$$
0.8 \underbrace{\begin{bmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 0 & 0 \\ 0 & 1/2 & 1 \end{bmatrix}}_{\mathbf{S}} + 0.2 \underbrace{\begin{bmatrix} 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \end{bmatrix}}_{1/n \cdot \mathbf{1} \cdot \mathbf{1}^T}
$$

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

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

Analytics and Visualization of Big Data

Fadel M. Megahed

Lecture 20: Link Analysis (Cont.)

AUBURN UNIVERSITY
SAMUEL GINN
COLLEGE OF ENGINEERING