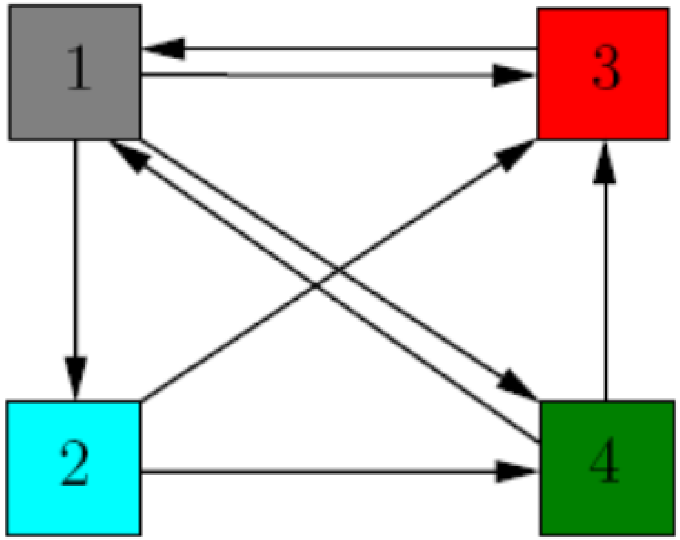
Link Analysis: PageRank Tutorial

Lin Guosheng
School of Computer Science and Engineering
Nanyang Technological University

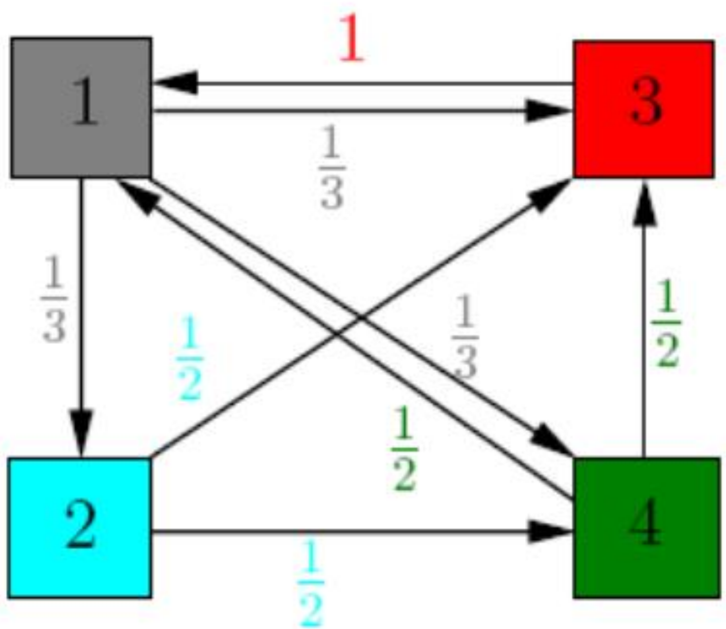
- Q1. Given the graph below, calculate the page rank score for each node using the following two methods:
  - 1) directly solving the flow equations;
  - 2) using the power iterative method for 2 iterations.



Answer:
Step 1: generate edge weights

- Define the edge weights as $\frac{1}{d_i}$.

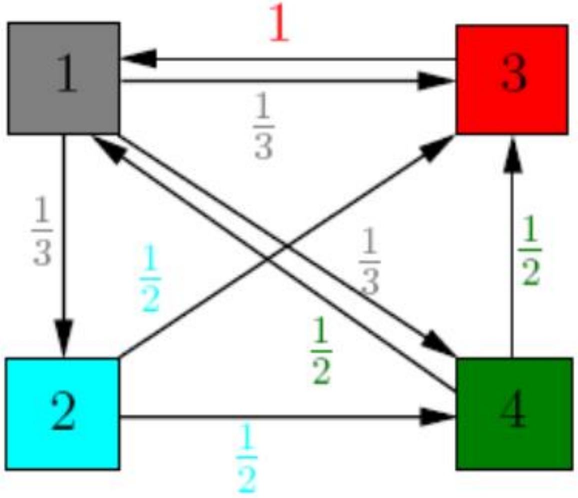
d: out-degree (the number of out-links)



We can go through the node one by one to generate the edge weights.
Example:
The out-links of node 1 is (1,3), (1,4) and (1,2)
For node 1, there are 3 out-links, so d=3.
The weight for each out-link of node1 is 1/3.

- Construct transition matrix M:

$$\mathbf{M} = \begin{array}{c} \begin{array}{cccc} \text{Node1} & \text{Node2} & \text{Node3} & \text{Node4} \end{array} \\ \begin{bmatrix} 0 & 0 & 1 & \frac{1}{2} \\ \frac{1}{3} & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{3} & \frac{1}{2} & 0 & 0 \end{bmatrix} \end{array}$$



Construct the transition matrix column by column
Each column corresponds to the out-link weights

Solution 1: solve flow equations

Solve flow equations for the PageRank scores: r

$$\mathbf{r} = [r_1,\ r_2,\ r_3,\ r_4]^\top$$

Flow equations (matrix expression):

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

$$\sum_i r_i = 1$$

$$\mathbf{M} = \begin{bmatrix} 0 & 0 & 1 & \frac{1}{2} \\ \frac{1}{3} & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{3} & \frac{1}{2} & 0 & 0 \end{bmatrix}$$

Flow equations:

$$r_1 = 1 \times r_3 + \frac{1}{2} \times r_4;$$

$$r_2 = \frac{1}{3} \times r_1;$$

$$r_3 = \frac{1}{3} \times r_1 + \frac{1}{2} \times r_2 + \frac{1}{2} \times r_4;$$

$$r_4 = \frac{1}{3} \times r_1 + \frac{1}{2} \times r_2;$$

$$r_1 + r_2 + r_3 + r_4 = 1$$

Solutions for scores r:

$$\frac{1}{31} \cdot \begin{bmatrix} 12 \\ 4 \\ 9 \\ 6 \end{bmatrix} \sim \begin{bmatrix} 0.38 \\ 0.12 \\ 0.29 \\ 0.19 \end{bmatrix}$$

- How to solve linear equations?
  - The simple method: repeatedly eliminate variables.
  - Other materials:
    - Refer to the linear algebra course SC1004/ CE1104/CZ1104
    - https://en.wikipedia.org/wiki/System_of_linear_equations

EQ5

Variable elimination

EQ1 $r_1 = 1 \times r_3 + \frac{1}{2} \times r_4;$

EQ2 $r_2 = \frac{1}{3} \times r_1;$

EQ3 $r_3 = \frac{1}{3} \times r_1 + \frac{1}{2} \times r_2 + \frac{1}{2} \times r_4;$

EQ4 $r_4 = \frac{1}{3} \times r_1 + \frac{1}{2} \times r_2;$

EQ5 $r_1 + r_2 + r_3 + r_4 = 1$

Variable elimination:
Here we aim to use r1 to express other variables

Substitute EQ2 into EQ4 (eliminate r2):
r4= 1/3 * r1 + (1/2) * (1/3) * r1
-> r4= 3/6 *r1 = 1/2 * r1 (EQ6)

Substitute EQ6 into EQ1 (eliminate r4):
r1= r3 + (1/2) * (1/2) * r1
-> r3= 3/4 * r1 (EQ7)

With EQ2, EQ6, EQ7 (eliminate r2,r4,r3),
We update EQ5:
r1 + 1/3 * r1 + 3/4 * r1 + 1/2 * r1= 1
-> (1 + 4/12 + 9/12 + 6/12) * r1 =1
-> 31/12 * r1 =1
-> r1 = 12/31

r1 = 12/31
r2 = 1/3 * r1 = 4/31 (from EQ2)
r3 = 3/4 * r1 = 9/31 (from EQ7)
r4 = 1/2 * r1= 6/31 (from EQ6)

Solution 2: Power iteration

- Power iteration method
  - 1. assign initial values to rank scores : r_i= 1/N;
    - N is the total number of nodes
  - 2. repeat the following util converge:

    Calculate the page rank:

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

Or equivalently for each node: $r_j^{(t+1)} = \sum_{i \to j} \frac{r_i^{(t)}}{d_i}$

Converge criteria: $(\sum_i \left| r_i^{t+1} - r_i^t \right| < \epsilon)$

Solution 2: Power iteration

$$\mathbf{r} = [r_1,\ r_2,\ r_3,\ r_4]^\top$$

$$\mathbf{M} = \begin{bmatrix} 0 & 0 & 1 & \frac{1}{2} \\ \frac{1}{3} & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{3} & \frac{1}{2} & 0 & 0 \end{bmatrix}$$

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

Iter 0: initialization:

$$\mathbf{r}^{(0)} = \begin{pmatrix} 0.25 \\ 0.25 \\ 0.25 \\ 0.25 \end{pmatrix}$$

Iter 1:

$$\mathbf{r}^{(1)} = \mathbf{M} \cdot \mathbf{r}^{(0)} = \begin{pmatrix} 0.37 \\ 0.08 \\ 0.33 \\ 0.20 \end{pmatrix}$$

Solution 2: Power iteration

$$\mathbf{r} = [r_1,\ r_2,\ r_3,\ r_4]^\top$$

$$\mathbf{M} = \begin{bmatrix} 0 & 0 & 1 & \frac{1}{2} \\ \frac{1}{3} & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{3} & \frac{1}{2} & 0 & 0 \end{bmatrix}$$

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

Iter 2:

$$\mathbf{r}^{(2)} = \mathbf{M} \cdot \mathbf{r}^{(1)} = \begin{pmatrix} 0.43 \\ 0.12 \\ 0.27 \\ 0.16 \end{pmatrix}$$

The question only requests to run 2 iterations.
For demonstration, here we show more iterations:

$$\mathbf{r}^{(3)} = \mathbf{M} \cdot \mathbf{r}^{(2)} = \begin{pmatrix} 0.35 \\ 0.14 \\ 0.29 \\ 0.20 \end{pmatrix}$$

The question only requests to run 2 iterations.
For demonstration, here we show more iterations:

Iter 4: $\begin{pmatrix} 0.39 \\ 0.11 \\ 0.29 \\ 0.19 \end{pmatrix}$

Iter 5: $\begin{pmatrix} 0.39 \\ 0.13 \\ 0.28 \\ 0.19 \end{pmatrix}$

Iter 6: $\begin{pmatrix} 0.38 \\ 0.13 \\ 0.29 \\ 0.19 \end{pmatrix}$

Iter 7 $\begin{pmatrix} 0.38 \\ 0.12 \\ 0.29 \\ 0.19 \end{pmatrix}$

Iter 8: $\begin{pmatrix} 0.38 \\ 0.12 \\ 0.29 \\ 0.19 \end{pmatrix}$

(converged, no significant changes)

Recap of Page Rank

- PageRank: Assign a rank score (importance score) to each node in the directed graph.
  - For a node j in the directed graph, we define page rank score r_i as follows (flow equations):

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

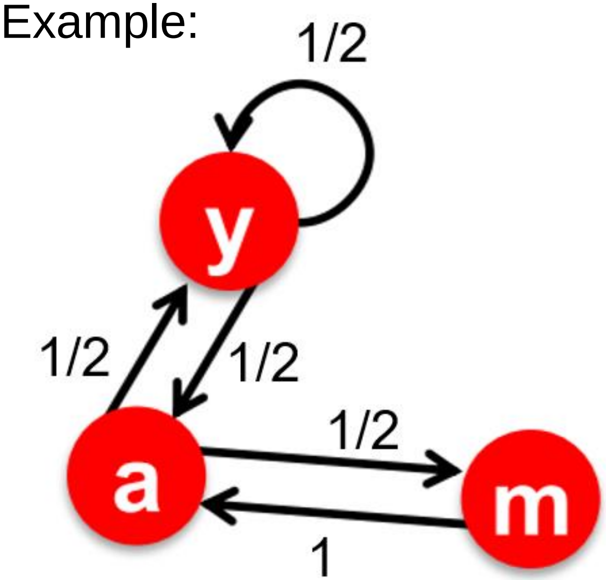
d_i … **out-degree of node** i

One node score = weighted sum of the neighbour node scores from in-links

the edge weights: $\frac{1}{d_i}$

We also require: $\sum_i r_i = 1$
(The sum of all rank scores equals 1):

Example:



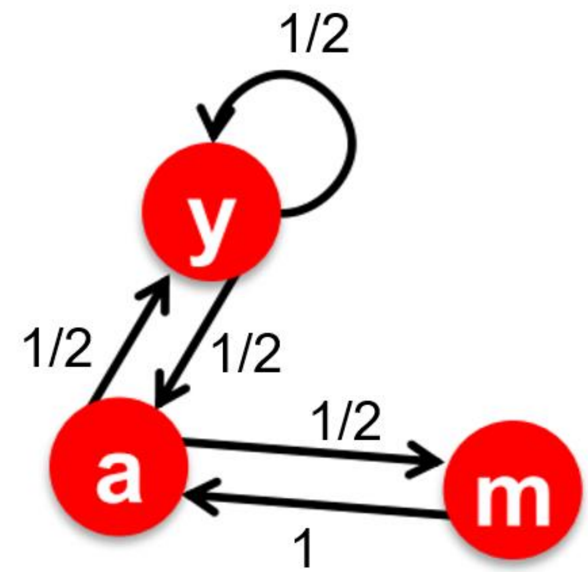
$r_y = r_y/2 + r_a/2$

Flow equation:
flow-out value = flow-in value

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

d_i … **out-degree of node** i

Another example:



"Flow" equations:

$r_y = r_y/2 + r_a/2$

$r_a = r_y/2 + r_m$

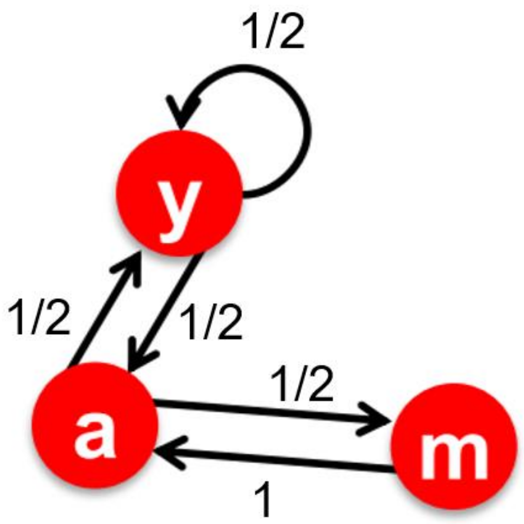
$r_m = r_a/2$

1. Directly solve linear equations

- Solution 1:
  Directly solve the linear equations

"Flow" equations:

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



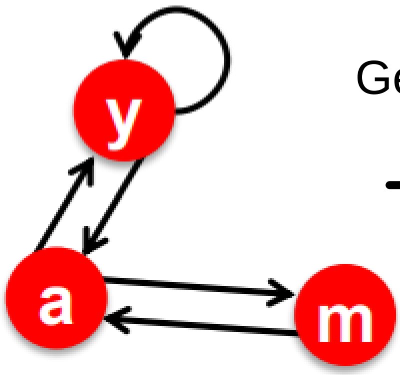
Adding this equation:

$$\sum_i r_i = 1$$

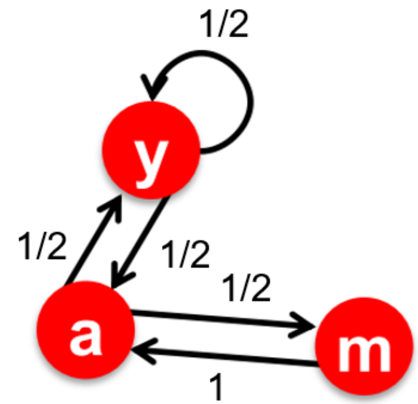
(The sum of all rank scores equals 1):

We can solve these 4 linear equations to get the solutions:
$r_y = 0.4$; $r_a = 0.4$; $r_m = 0.2$

Use transition matrix



Generate edge weights



Each column in M indicates out-links

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

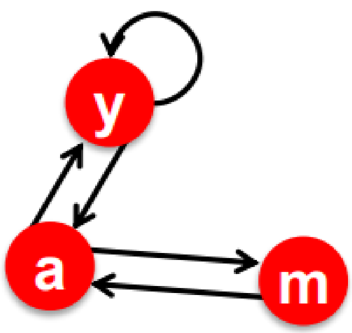
Transition matrix M

- Rank vector **r :** a column vector of all rank scores

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

Matrix expression of flow equations

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



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

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

Matrix expression $\longrightarrow$

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

2. Power iteration method

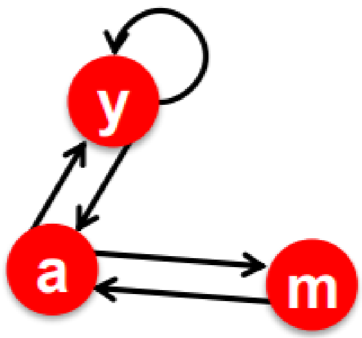
- Power iteration method
  - 1. assign initial values to rank scores : $r_i = 1/N$;
    - N is the total number of nodes
  - 2. repeat the following util converge:

    Calculate the page rank:

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

Or equivalently for each node: $r_j^{(t+1)} = \sum_{i \to j} \frac{r_i^{(t)}}{d_i}$

Converge criteria: $\left(\sum_i \left| r_i^{t+1} - r_i^t \right| < \epsilon\right)$



	y	a	m
y	½	½	0
a	½	0	1
m	0	½	0

$$
\begin{aligned}
r_y &= r_y/2 + r_a/2 \\
r_a &= r_y/2 + r_m \\
r_m &= r_a/2
\end{aligned}
$$

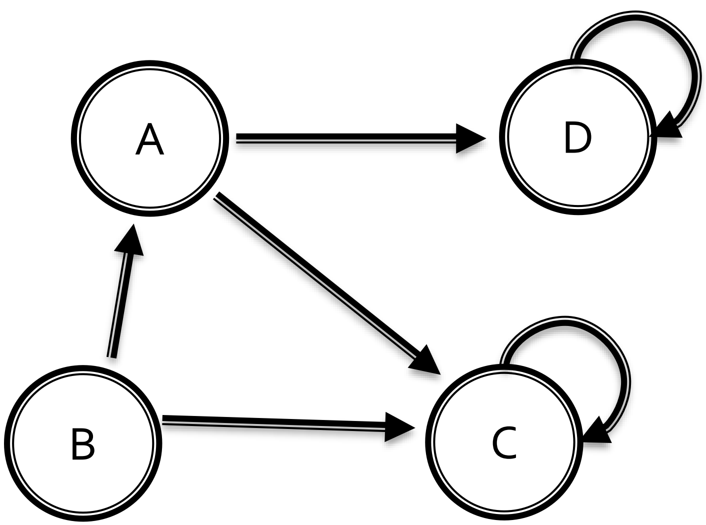
- **Power Iteration:**
  - Set $r_j \leftarrow 1/N$
  - **1:** $r'_j \leftarrow \sum_{i\to j} \frac{r_i}{d_i}$
  - **2:** If $|r - r'| > \varepsilon$:
    - $r \leftarrow r'$
  - **3:** go to **1**

- **Example:**

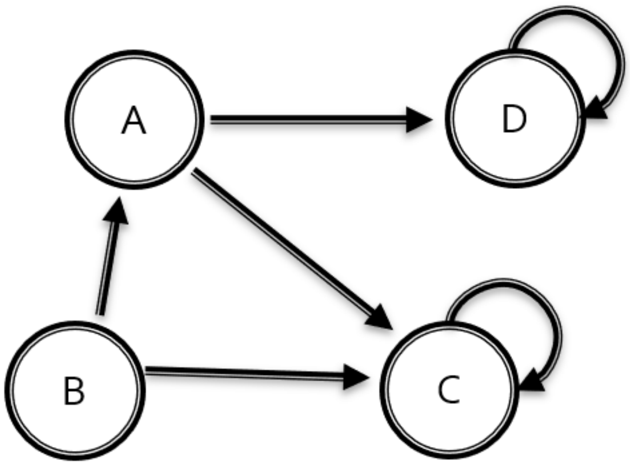
$$
\begin{pmatrix} r_y \\ r_a \\ r_m \end{pmatrix} = \begin{pmatrix} 1/3 \\ 1/3 \\ 1/3 \end{pmatrix} \begin{pmatrix} 1/3 \\ 3/6 \\ 1/6 \end{pmatrix} \begin{pmatrix} 5/12 \\ 1/3 \\ 3/12 \end{pmatrix} \begin{pmatrix} 9/24 \\ 11/24 \\ 1/6 \end{pmatrix} \ldots \begin{pmatrix} 6/15 \\ 6/15 \\ 3/15 \end{pmatrix}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

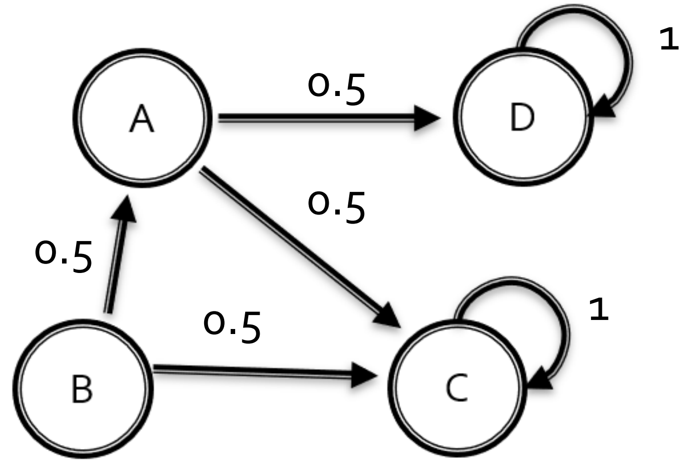
Initial value: 1/N (N=3) | iter1 | iter2 | iter3 | Converged

- Q2. A graph is given below.
  - a) calculate the page rank score for each node using the power iteration method for 3 iterations.
  - b) identify one spider trap group if there are any.



- Solution (a):





Transition matrix M

A	B	C	D
0	0.5	0	0
0	0	0	0
0.5	0.5	1	0
0.5	0	0	1

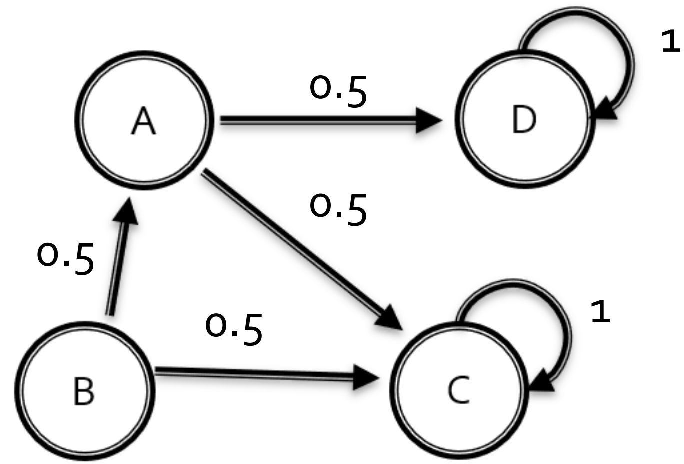
$$r^{(t+1)} = M \cdot r^{(t)}$$

Transition matrix M

0	0.5	0	0
0	0	0	0
0.5	0.5	1	0
0.5	0	0	1

Initial iteration

0.25
0.25
0.25
0.25



0.125
0
0.5
0.375

Iteration 1

0
0
0.5625
0.4375

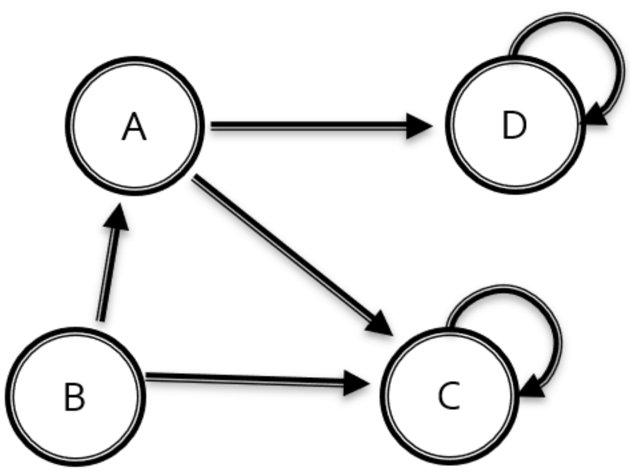
Iteration 2

0
0
0.5625
0.4375

Iteration 3

b) identify one spider trap group if there are any.

Solution (b):

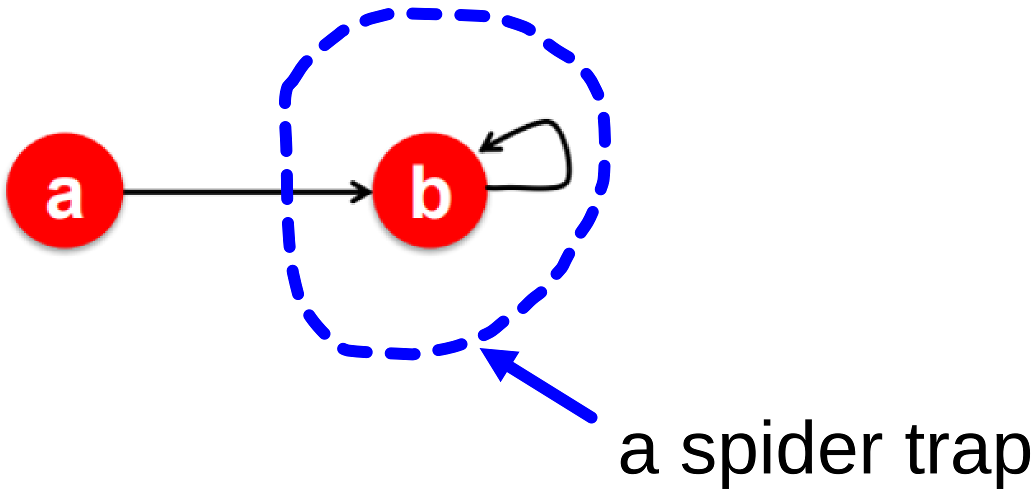


Spider trap groups:

1. {C}
2. {D}
3. {A, C, D}
4. {C, D}

(You only need to specify one group for this question)

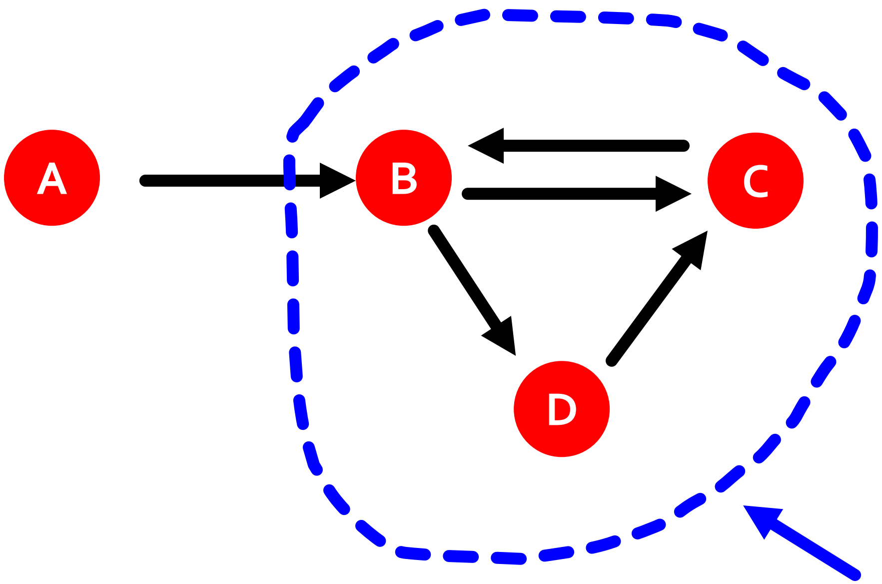
2. The spider trap problem:



A spider trap group: for the nodes in the group, all out-links are within the group

After entering the spider-trap group, the web user cannot go out of the group by navigating through the links

Another spider trap group example:
(for the nodes in the group, all out-links are within the group)



the group forms a spider trap problem

After entering the spider-trap group, the web user cannot go out of the group by navigating through the links