CSE 4/587

Data Intensive Computing

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

Day 11

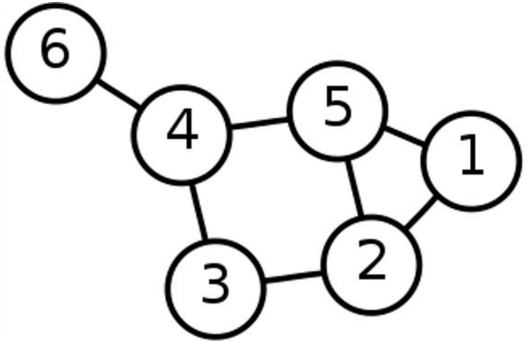
Graph Analytics and PageRank

Announcements and Feedback

- Project Phase 1 due Monday @ 11:59PM
    - Submission will be setup through UBLearns

What is a Graph?

- A ***graph*** is a structure made up of a set of objects, where some pairs of the objects are "related"
- Mathematically, objects are represented with ***vertices*** (or nodes or points) and the relations between two vertices are represented with ***edges*** (or links or lines)

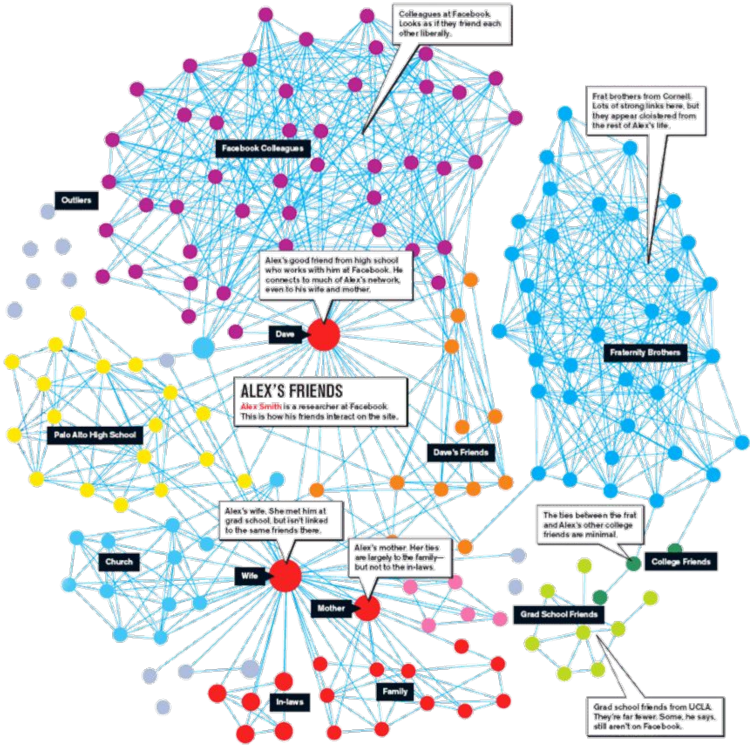


- Typically, a graph is depicted in diagrammatic form as a set of dots or circles for the vertices, joined by lines or curves for the edges
- Edges can be directed or undirected (a relationship can go both ways)
- Edges can be weighted to show the “strength”, distance, etc

Graph Structure is Everywhere

- Any social media application that you are a part of can be modeled as a graph
  - Vertices are users, posts or images
  - Edges are any social relationship between them.

ie: a person hitting 'like' on a particular post/image can be considered an edge



ALEX'S FRIENDS

Alex Smith is a researcher at Facebook. This is how his friends interact on the site.

Colleagues at Facebook. Looks as if they friend each other liberally.

Frat brothers from Cornell. Lots of strong links here, but they appear cloistered from the rest of Alex's life.

Alex's good friend from high school who works with him at Facebook. He connects to much of Alex's network, even to his wife and mother.

Alex's wife. She met him at grad school, but isn't linked to the same friends there.

Alex's mother. Her ties are largely to the family—but not to the in-laws.

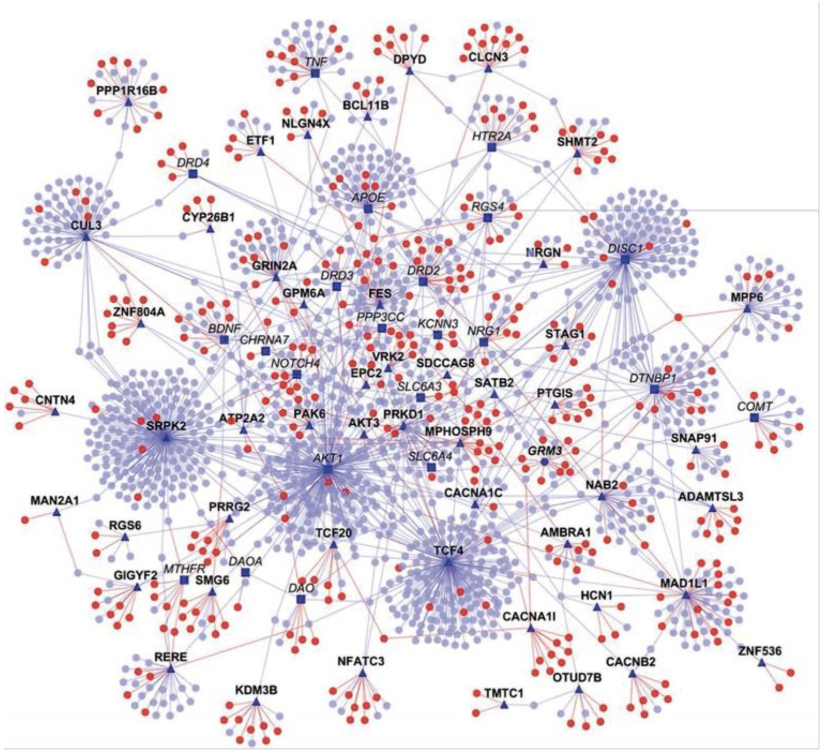
The ties between the frat and Alex's other college friends are minimal.

Grad school friends from UCLA. They're far fewer. Some, he says, still aren't on Facebook.

Ref :https://www.pinterest.com/pin/490470215639647556/

Graph Structure is Everywhere

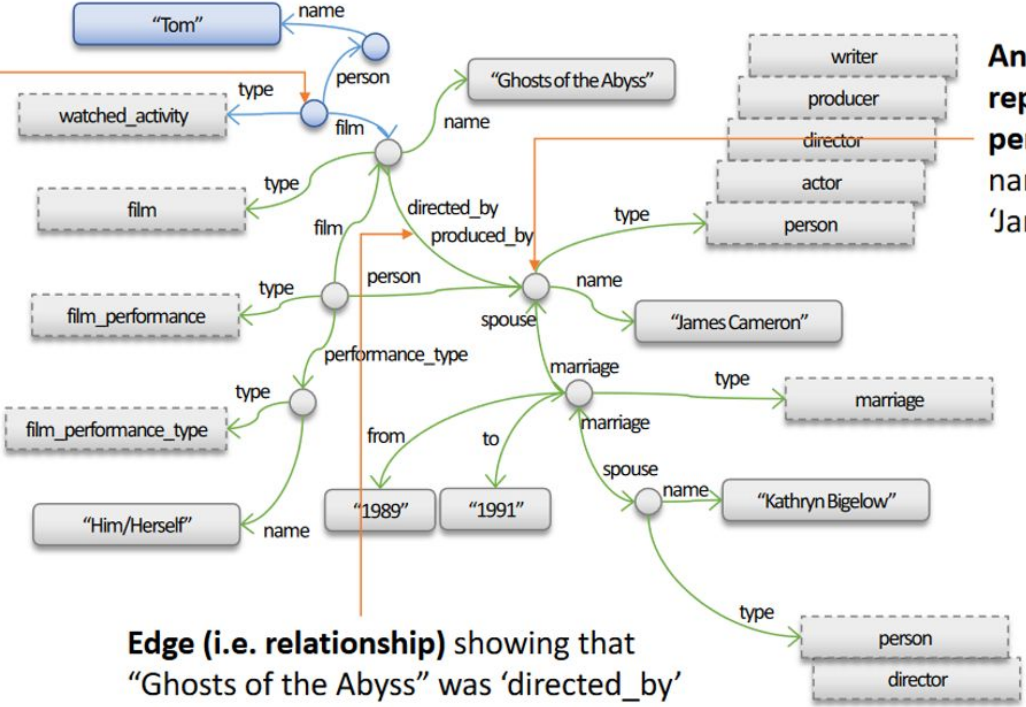
- A Biological network can be modelled as a graph
- For example, a Protein-Protein interaction graph
  - Vertices are proteins
  - Edges are interaction between them



Ref :https://www.genengnews.com/insights/protein-protein-interactions-get-a-new-groove-on/

Knowledge Graphs

Personal entity showing that Tom watched Ghosts of the Abyss



An entity representing a person with name **attribute** 'James Cameron'

Edge (i.e. relationship) showing that "Ghosts of the Abyss" was 'directed_by' and 'produced_by' James Cameron

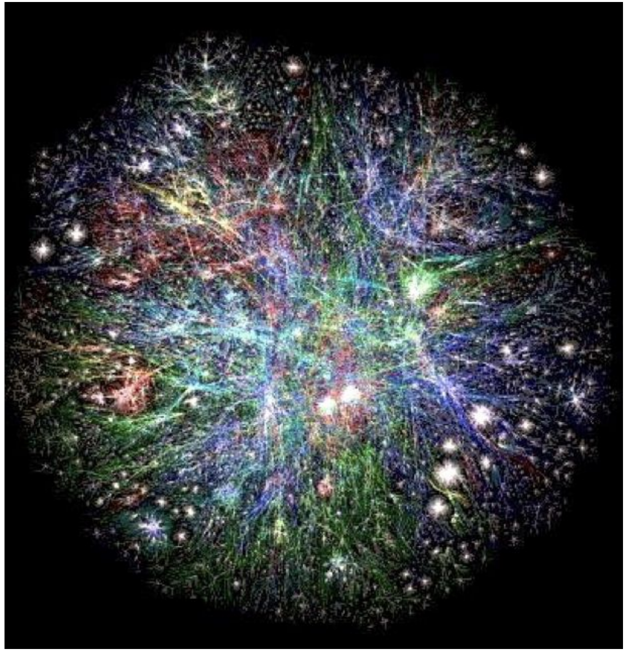
Key concepts

Entity	Represent something in the real world
Edge	Represent relationship
Attribute	Represent something about an entity
Ontology	Definition of possible types of entities, relationships and attributes

Graph Structure is Everywhere

- The internet can be modeled as a graph
  - Vertices are webpages
  - Edges are the links between them
- This modeling can be used to compute the importance of each webpage in the network

This will be the topic of the next few lectures



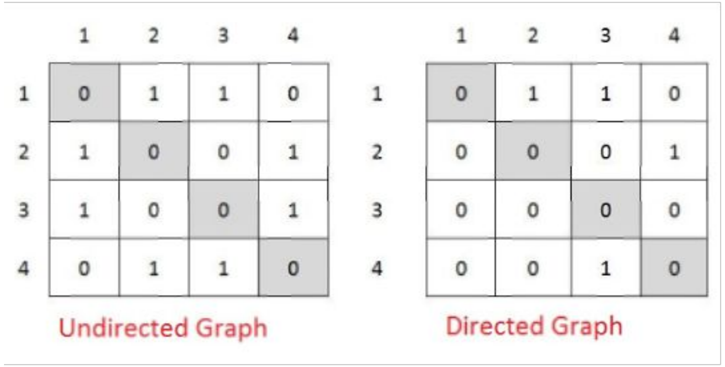
Ref : http://www.vlib.us/web/worldwideweb3d.html

Graph Representations

There are two standard ways to represent a graph G(V,E) *[V is the set of vertices, E is the set of edges]*

1. adjacency list representation
2. adjacency matrix

An adjacency matrix is 2-Dimensional Array of size *V*x*V*, where *V* is the number of vertices in the graph.



	1	2	3	4
1	0	1	1	0
2	1	0	0	1
3	1	0	0	1
4	0	1	1	0

Undirected Graph

	1	2	3	4
1	0	1	1	0
2	0	0	0	1
3	0	0	0	0
4	0	0	1	0

Directed Graph

An adjacency list is an array of linked lists, where the array size is same as number of vertices in the graph. Every vertex has a linked list. Each node in this linked list represents the reference to another vertex that shares an edge with the current vertex.

Broad Question

How can we organize the internet?

Broad Question

How can we organize the internet?

First try: Human Curated

- Web directories
- Yahoo, DMOZ, LookSmart

Broad Question

How can we organize the internet?

First try: Human Curated

- Web directories
- Yahoo, DMOZ, LookSmart

Second try: Web Search

- Information retrieval investigates to find relevant docs in a small and trusted set of newspaper articles, patents, etc.

Broad Question

How can we organize the internet?

First try: Human Curated

- Web directories
- Yahoo, DMOZ, LookSmart

Second try: Web Search

- Information retrieval investigates to find relevant docs in a small and trusted set of newspaper articles, patents, etc.

But: Web is huge, full of untrusted documents, random things, web spam, etc

Web Search Challenges

Web contains many sources of information…**which can we “trust”?**

Web Search Challenges

Web contains many sources of information…**which can we “trust”?**

If we know one trustworthy page, it may point to another.

Web Search Challenges

Web contains many sources of information…**which can we “trust”?**

If we know one trustworthy page, it may point to another.

What is the “best” answer to the query “newspaper”?

Web Search Challenges

Web contains many sources of information…**which can we “trust”?**

If we know one trustworthy page, it may point to another.

What is the “best” answer to the query “newspaper”?

- There is no one single right answer to the question
- Pages that actually know about newspapers might all be pointing to many different newspapers

Ranking Nodes in a Graph

All web pages are not equally “important”

Some websites may provide more trustworthy information

Consider the following websites:

www.joe-schmoe.com *vs* www.buffalo.edu *or* www.stanford.edu

Ranking Nodes in a Graph

All web pages are not equally “important”

Some websites may provide more trustworthy information

Consider the following websites:

www.joe-schmoe.com *vs* www.buffalo.edu *or* www.stanford.edu

The university websites are more important than the other website

Ranking Nodes in a Graph

All web pages are not equally “important”

Some websites may provide more trustworthy information

Consider the following websites:

www.joe-schmoe.com *vs* www.buffalo.edu *or* www.stanford.edu

The university websites are more important than the other website

So how do we come up with a system to rank this pages?

Link Analysis Algorithm

Key idea is to use links between pages as ***votes***

A page is more important if it has more links associated with it

What kind of links are more important? Incoming or outgoing?

Link Analysis Algorithm

Key idea is to use links between pages as ***votes***

A page is more important if it has more links associated with it

What kind of links are more important? Incoming or outgoing?

The incoming links are more important!

www.buffalo.edu is referred to in lot of other pages. So it must be a pretty influential page.

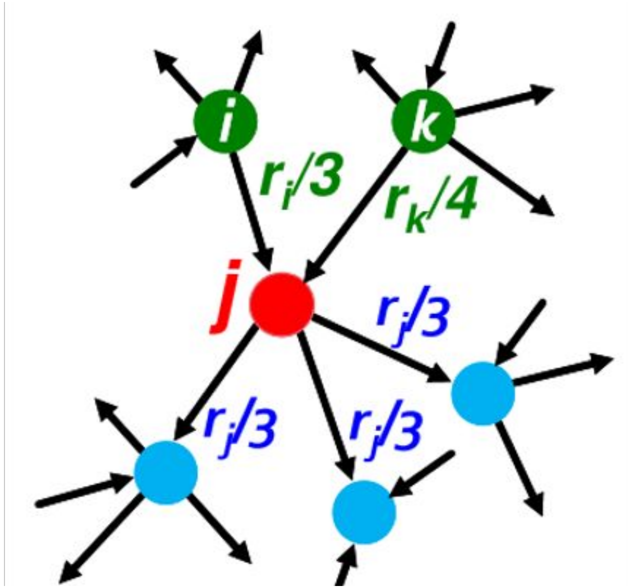
So do all incoming links have equal weightage?

Recursive Formulation

Each link's vote is proportional to the importance of its source page

If page ***j*** with importance $\boldsymbol{r_j}$ has ***n*** out-links, each link gets $\boldsymbol{r_j / n}$ votes

Page ***j***'s own importance is the sum of the votes on its in-links



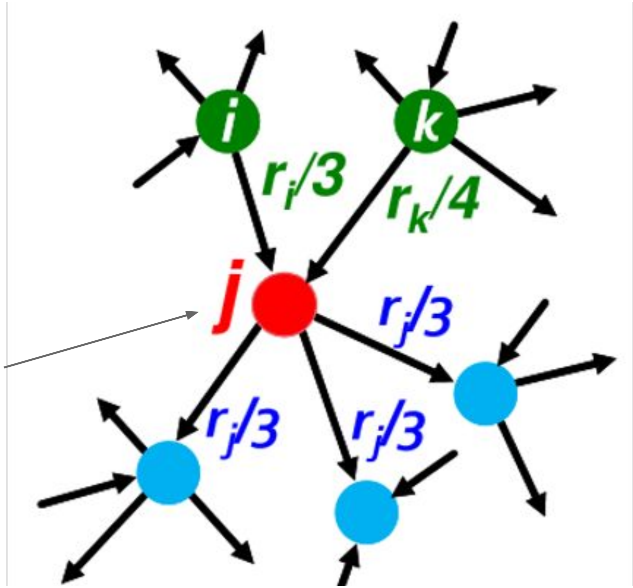
Recursive Formulation

Each link's vote is proportional to the importance of its source page

If page ***j*** with importance $\boldsymbol{r_j}$ has ***n*** out-links, each link gets $\boldsymbol{r_j / n}$ votes

Page ***j***'s own importance is the sum of the votes on its in-links

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



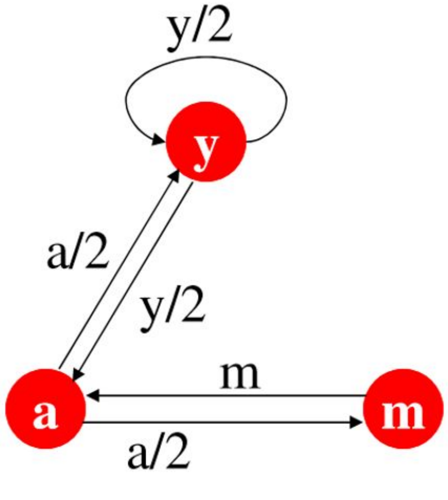
Page Rank: The Flow Model

A link from an *important page* (higher ranking page) is worth more

A page is *important* if it is pointed to by other important pages

Define a “rank” $\boldsymbol{r_j}$ for page ***j*** as:

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



“Flow” equations:

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

Solving the Flow Equation

3 equations, 3 unknowns, no constants

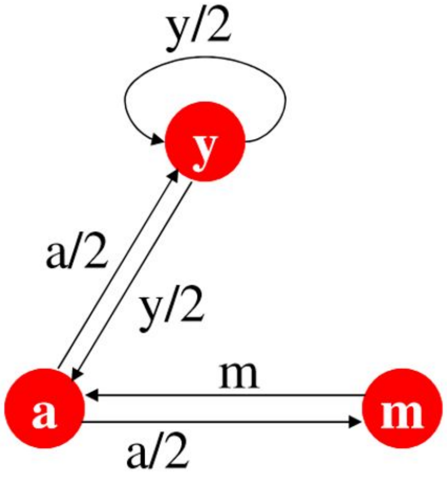
No unique solution: All solutions equivalent modulo the scale factor

Adding an additional constraint forces uniqueness:

$$r_y + r_a + r_m = 1$$

Gaussian Elimination can be used to find the solution.

This method will work for small graphs, but won't scale for larger graphs



"Flow" equations:

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

$$r_a = r_y/2 + r_m$$

$$r_m = r_a/2$$

Page Rank: Matrix Formulation

Stochastic Adjacency matrix $\boldsymbol{M}$

$\boldsymbol{M_{ji} = 1/(d_i)}$ if there is a link from $\boldsymbol{i}$ to $\boldsymbol{j}$, else value is 0

Page Rank: Matrix Formulation

Stochastic Adjacency matrix ***M***

$\boldsymbol{M_{ji}} = \boldsymbol{1/(d_i)}$ if there is a link from ***i*** to ***j***, else value is 0

If ***r*** is vector with the initial importance of a page and

$$\Sigma_i r_i \quad = 1$$

Page Rank: Matrix Formulation

Stochastic Adjacency matrix ***M***

$\boldsymbol{M_{ji}} = \boldsymbol{1/(d_i)}$ if there is a link from ***i*** to ***j***, else value is 0

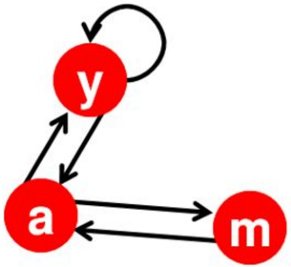
If ***r*** is vector with the initial importance of a page and

$$\sum_i r_i = 1$$

Then the flow equation can be written as

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

Solving with Power Iteration



	y	a	m
y	½	½	0
a	½	0	1
m	0	½	0

$$r = M \cdot r$$

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

$$r_a = r_y/2 + r_m$$

$$r_m = r_a/2$$

$$\begin{bmatrix} y \\ a \\ m \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{2} & 0 & 1 \\ 0 & \frac{1}{2} & 0 \end{bmatrix} \begin{bmatrix} y \\ a \\ m \end{bmatrix}$$

Solving with Power Iteration

Given a web graph with ***n*** nodes, where the vertices are pages and edges are hyperlinks

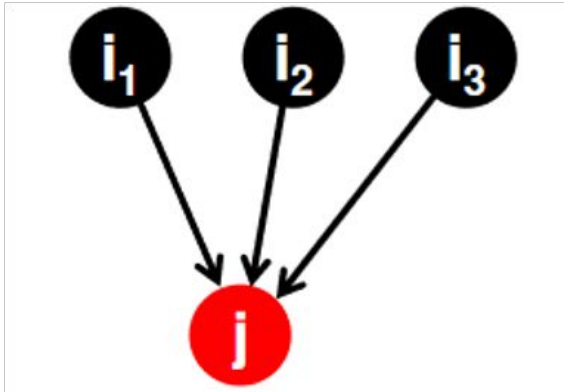
Power iteration: a simple iterative scheme

Suppose there are ***N*** web pages

1. **Initialize:** $r(0) = [1/N,\ldots.,1/N]^T$
2. **Iterate:** $r(t+1) = M \cdot r(t)$
3. **Stop when:** $||r(t+1) - r(t)||_1 < \varepsilon$

Random Walk Interpretation

Imagine a random web surfer

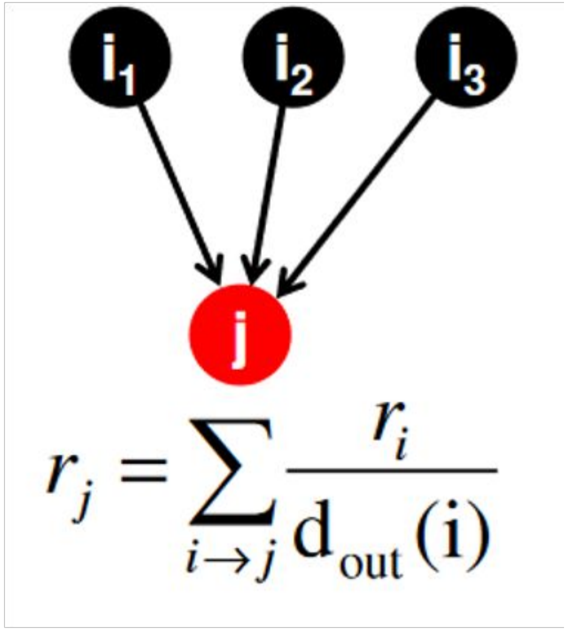


$$r_j = \sum_{i \to j} \frac{r_i}{\mathrm{d}_{\mathrm{out}}(\mathrm{i})}$$

Random Walk Interpretation

Imagine a random web surfer

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

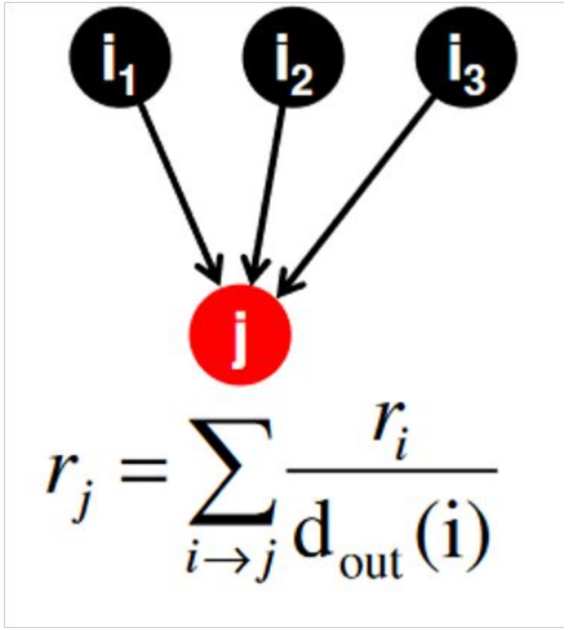


$$r_j = \sum_{i \to j} \frac{r_i}{\mathrm{d}_{\mathrm{out}}(\mathrm{i})}$$

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***

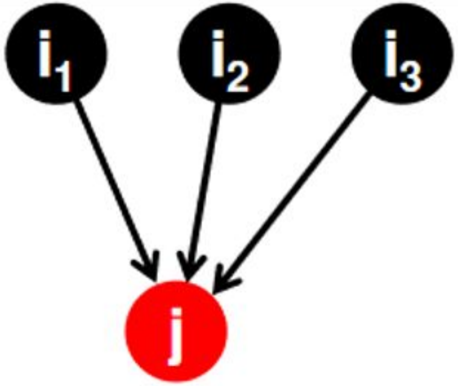


$$r_j = \sum_{i \to j} \frac{r_i}{\mathrm{d}_{\mathrm{out}}(\mathrm{i})}$$

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



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

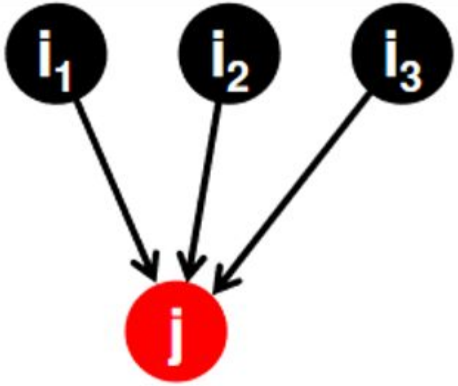
Random Walk Interpretation

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

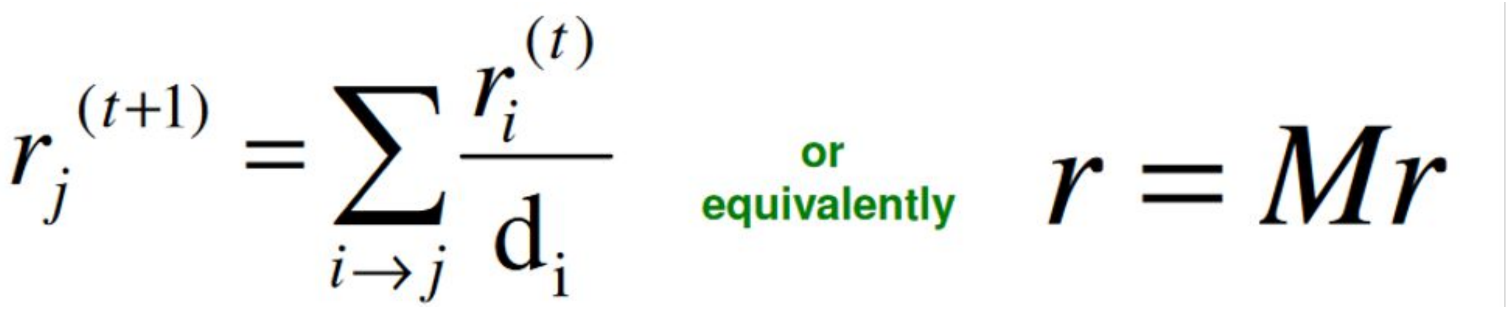
P*(*t*)** is the vector whose ***ith coordinate is the probability that the surfer is at page ***i*** at time ***t***

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



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

Google Formulation



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

or equivalently

$$r = Mr$$

Google Formulation

$$r_j^{(t+1)} = \sum_{i \to j} \frac{r_i^{(t)}}{\mathrm{d}_\mathrm{i}} \quad \text{or equivalently} \quad r = Mr$$

Does this value converge ?

Google Formulation

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

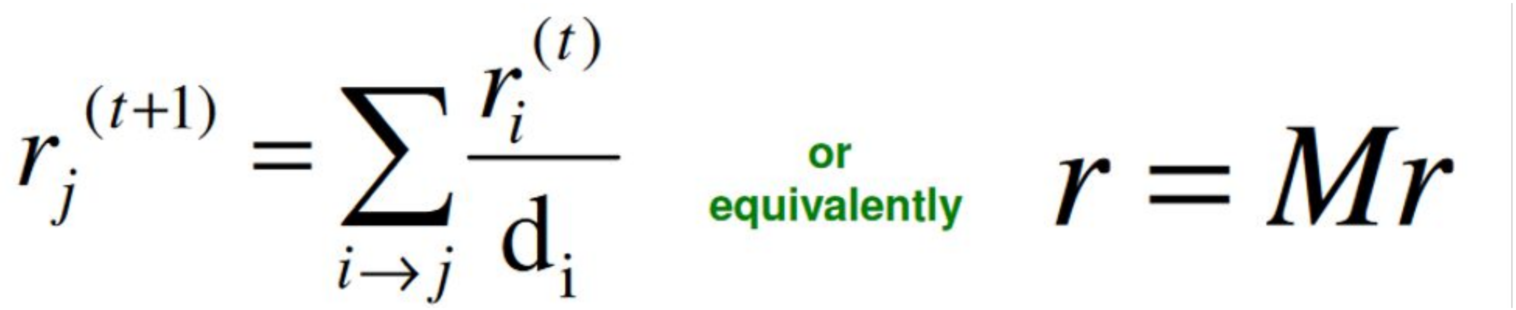
or equivalently

$$r = Mr$$

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

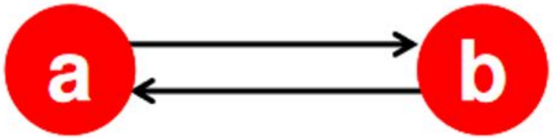
$$r = Mr$$

Does this value converge ?

Does it converge to the results that we want?

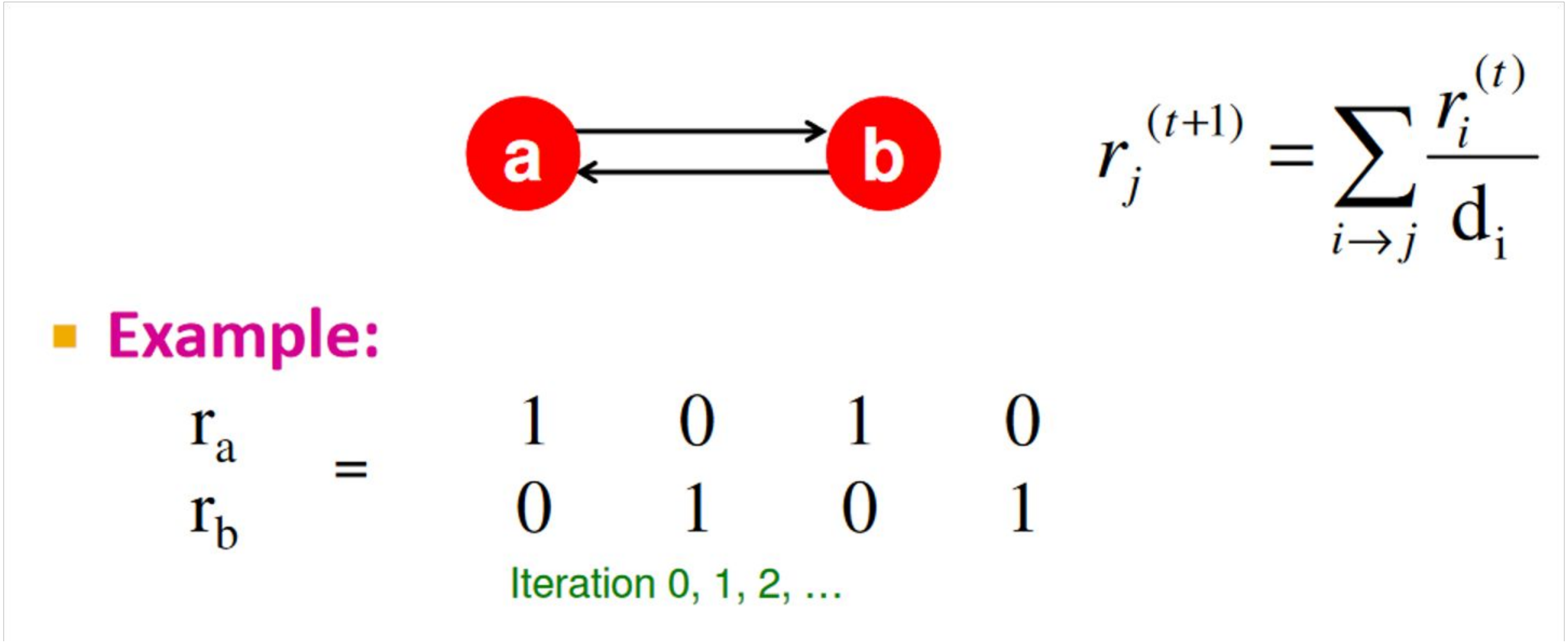
Are the results reasonable?

Does this converge?



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

Does this converge?



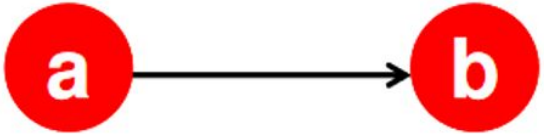
$$r_j^{(t+1)} = \sum_{i \to j} \frac{r_i^{(t)}}{\mathrm{d}_\mathrm{i}}$$

- **Example:**

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

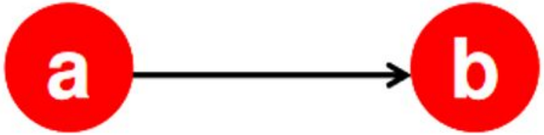
Iteration 0, 1, 2, …

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?



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

- **Example:**

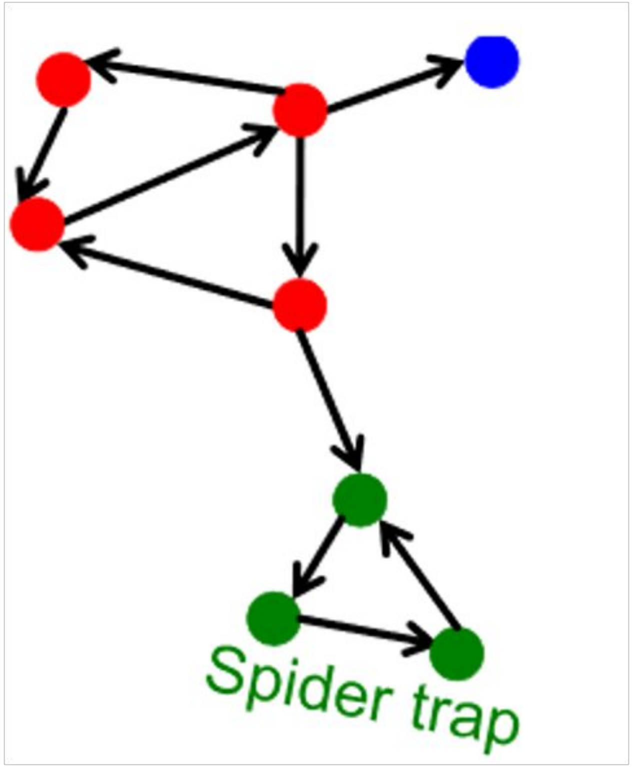
$$\begin{array}{c} \mathrm{r_a} \\ \mathrm{r_b} \end{array} = \begin{array}{cccc} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{array}$$

Iteration 0, 1, 2, …

Page Rank: Problems

Some pages are dead ends:

- Random walk has nowhere to go
- Such pages cause important information to leak



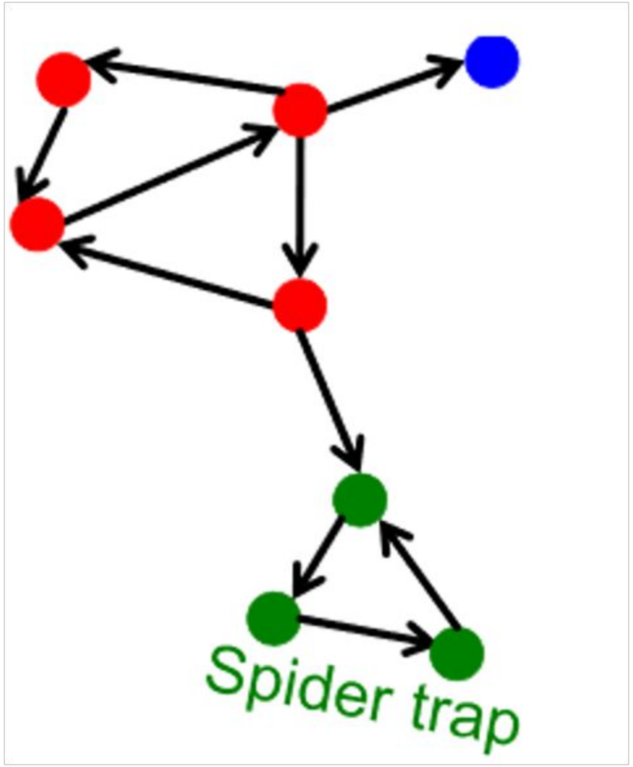
Page Rank: Problems

Some pages are dead ends:

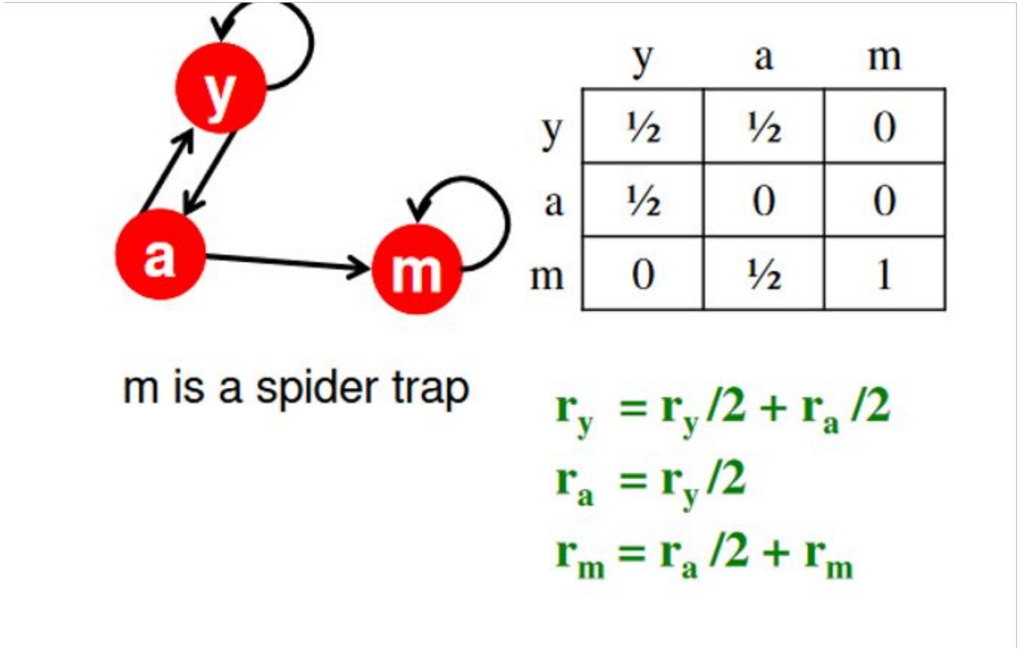
- Random walk has nowhere to go
- Such pages cause important information to leak

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

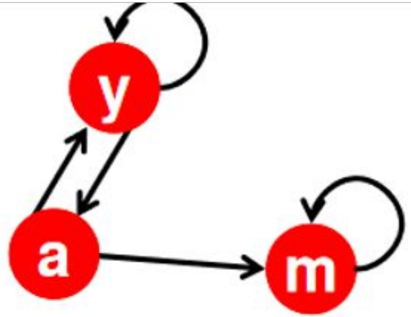
	y	a	m
y	½	½	0
a	½	0	0
m	0	½	1

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

$$r_a = r_y/2$$

$$r_m = r_a/2 + r_m$$

Spider Traps



$$
\begin{array}{c|c|c|c|}
 & y & a & m \\
\hline
y & ½ & ½ & 0 \\
\hline
a & ½ & 0 & 0 \\
\hline
m & 0 & ½ & 1 \\
\hline
\end{array}
$$

m is a spider trap

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

Iteration 0, 1, 2, …

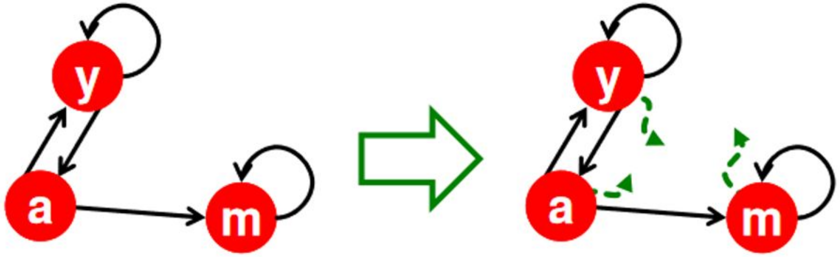
All the PageRank score gets "trapped" in node m.

Solution: Teleports

The Google solution for spider traps: Teleports

At each time step, the random surfer has two options:

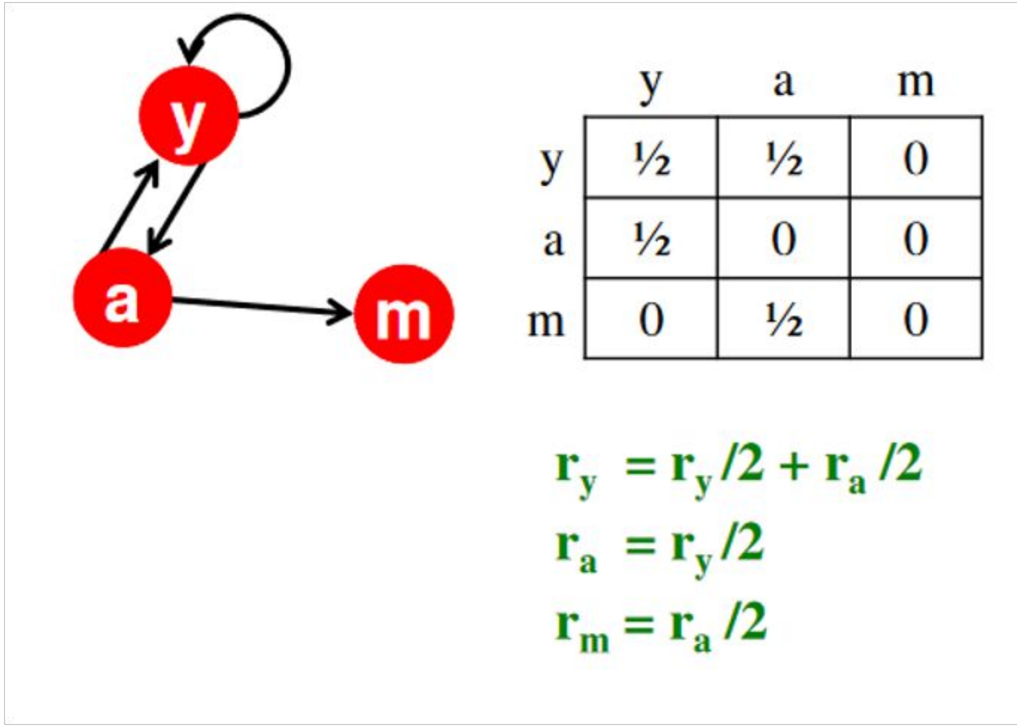
1. With probability β, follow a link at random
2. With prob. 1-β, jump to some random page



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



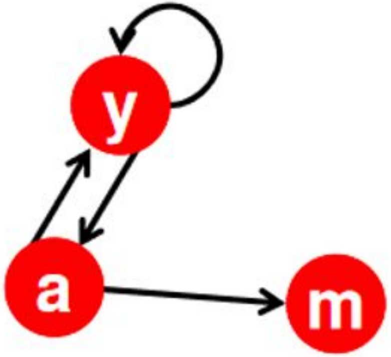
	y	a	m
y	½	½	0
a	½	0	0
m	0	½	0

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

$$r_a = r_y/2$$

$$r_m = r_a/2$$

Dead Ends



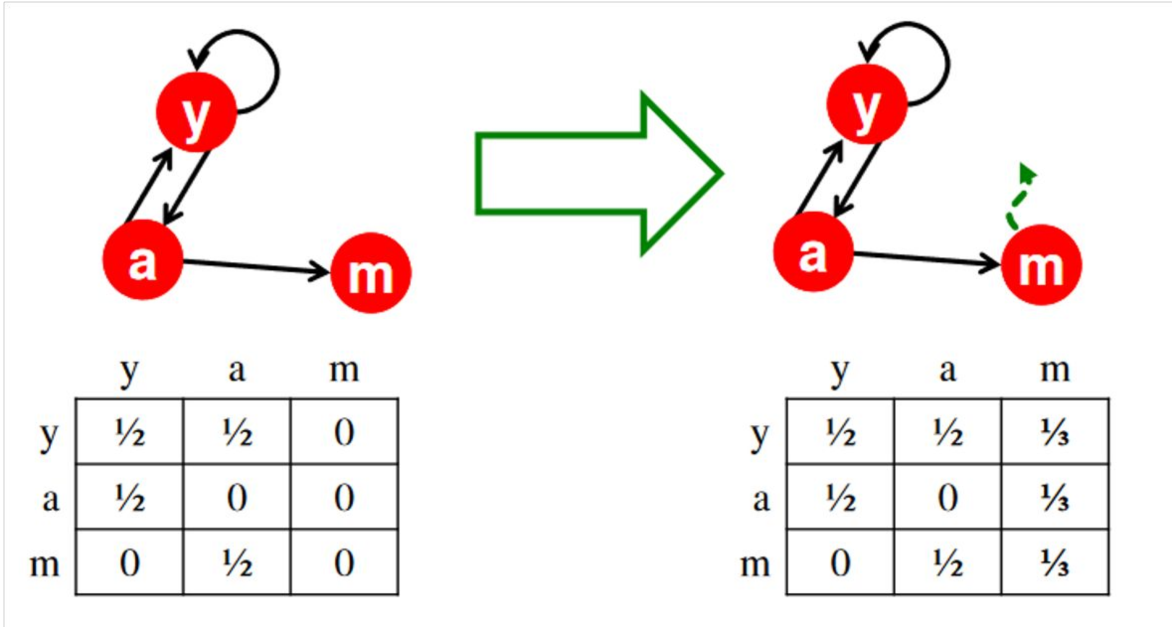
	y	a	m
y	½	½	0
a	½	0	0
m	0	½	0

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

Iteration 0, 1, 2, …

Solution: Teleports

Teleport with probability 1.0 at dead ends



	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

Google's Solution

Googles solution for PageRank:

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

Google's Solution

Googles solution for PageRank:

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

In matrix notation:

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

References

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