

# Chapter 8: Graph Data

## Part 1: Link Analysis & Page Rank

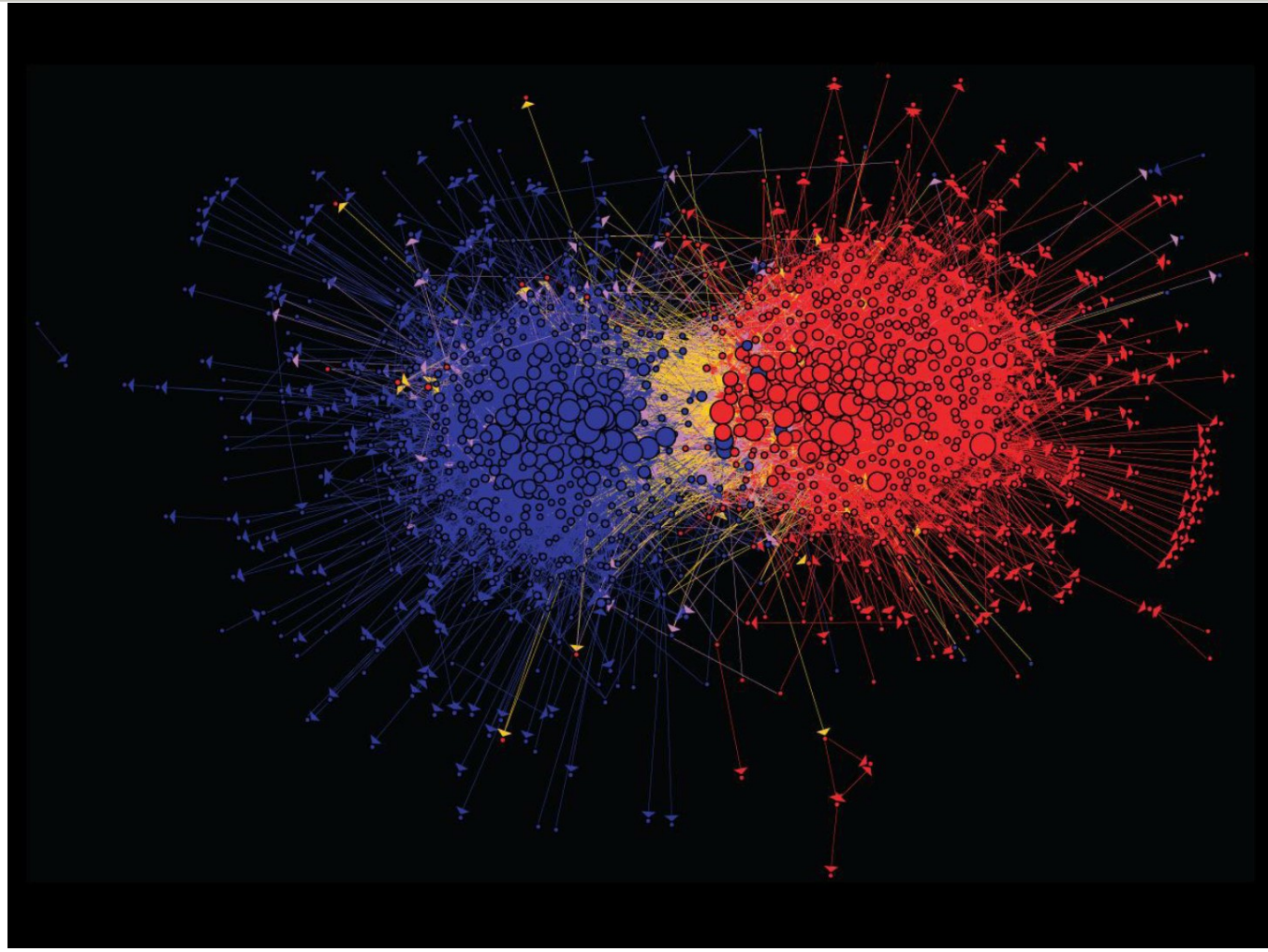
Based on  
Leskovec, Rajaraman, Ullman 2014:  
Mining of Massive Datasets



facebook

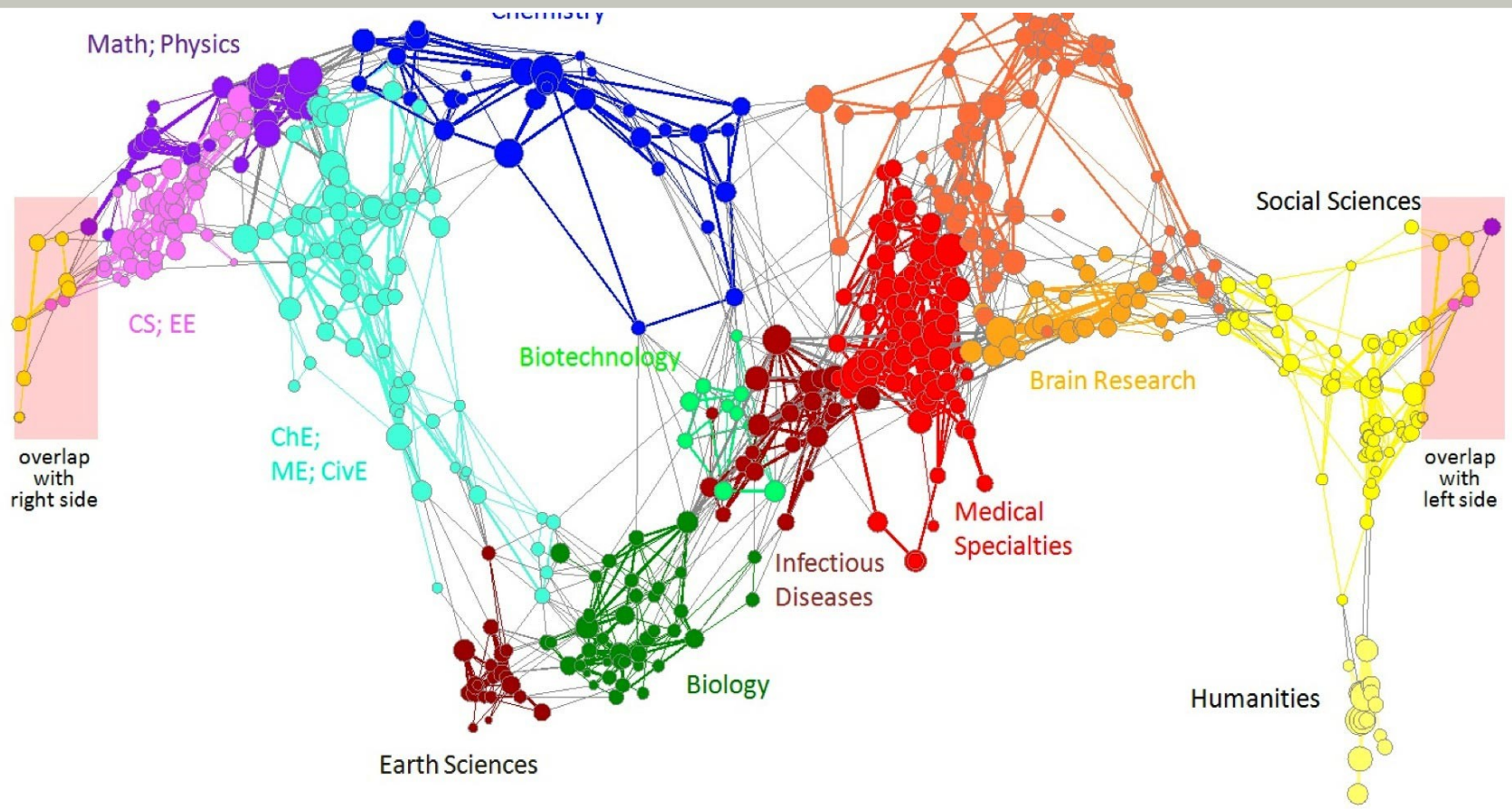
[Source: 4-degrees of separation, Backstrom-Boldi-Rosa-Ugander-Vigna, 2011]

December 2010



Connections between political blogs  
Polarization of the network [Adamic-Glance, 2005]

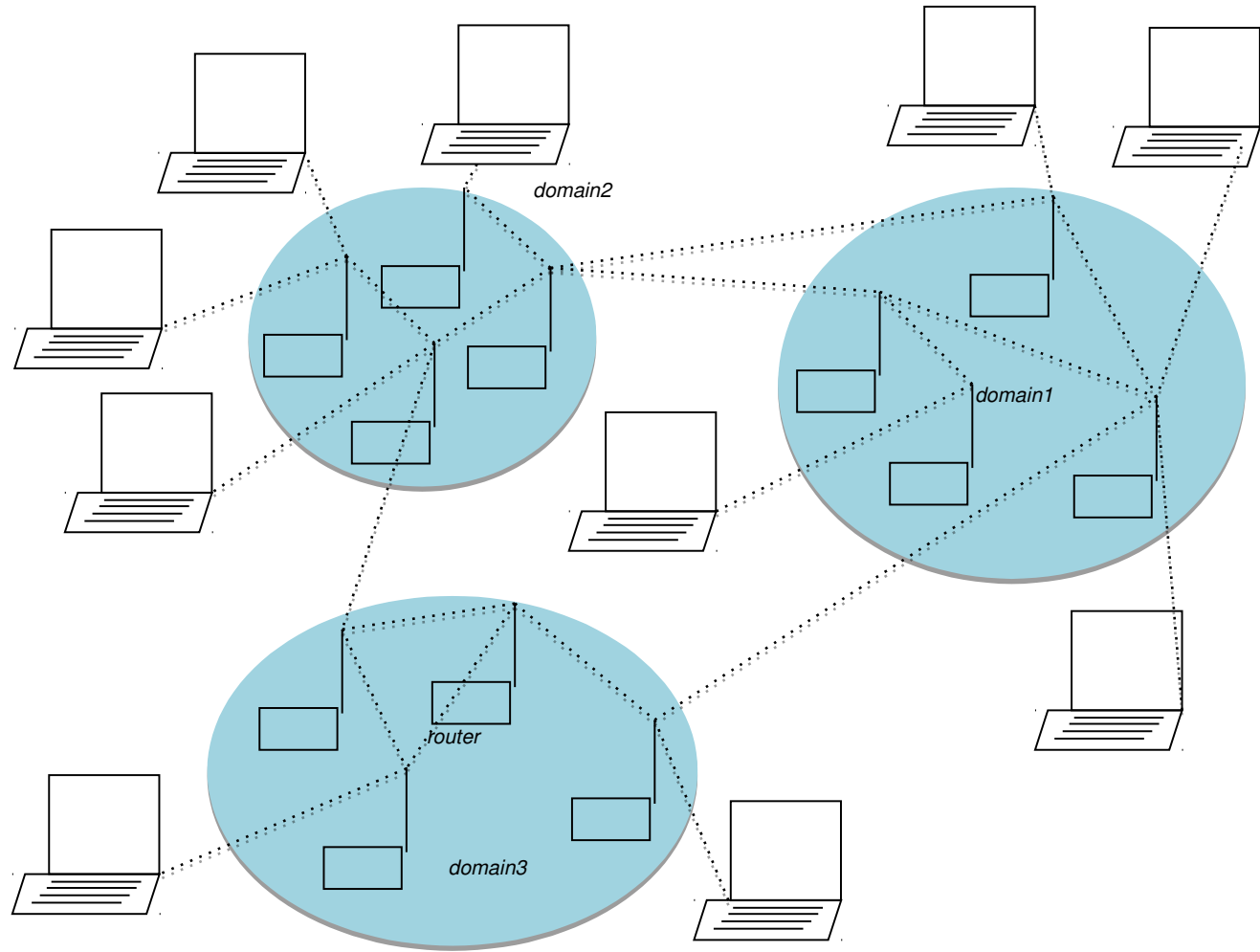
# Graph Data: Information Networks



Citation Networks and Map of Science  
[Börner et al., 2012]



Road Network of Toulouse  
[Mathieu Leplatre]



The Internet

## Web as a directed graph:

- Nodes: Webpages
- Edges: Hyperlinks

LMU  
München

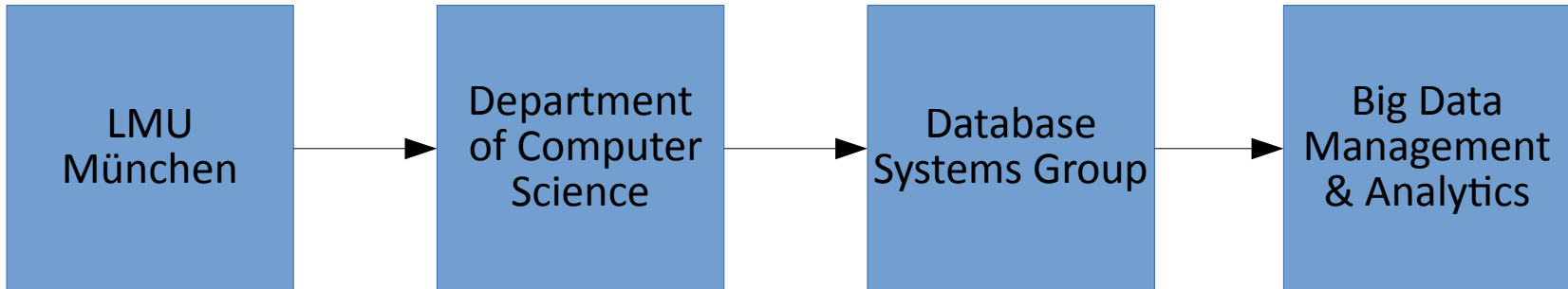
Department  
of Computer  
Science

Database  
Systems Group

Big Data  
Management  
& Analytics

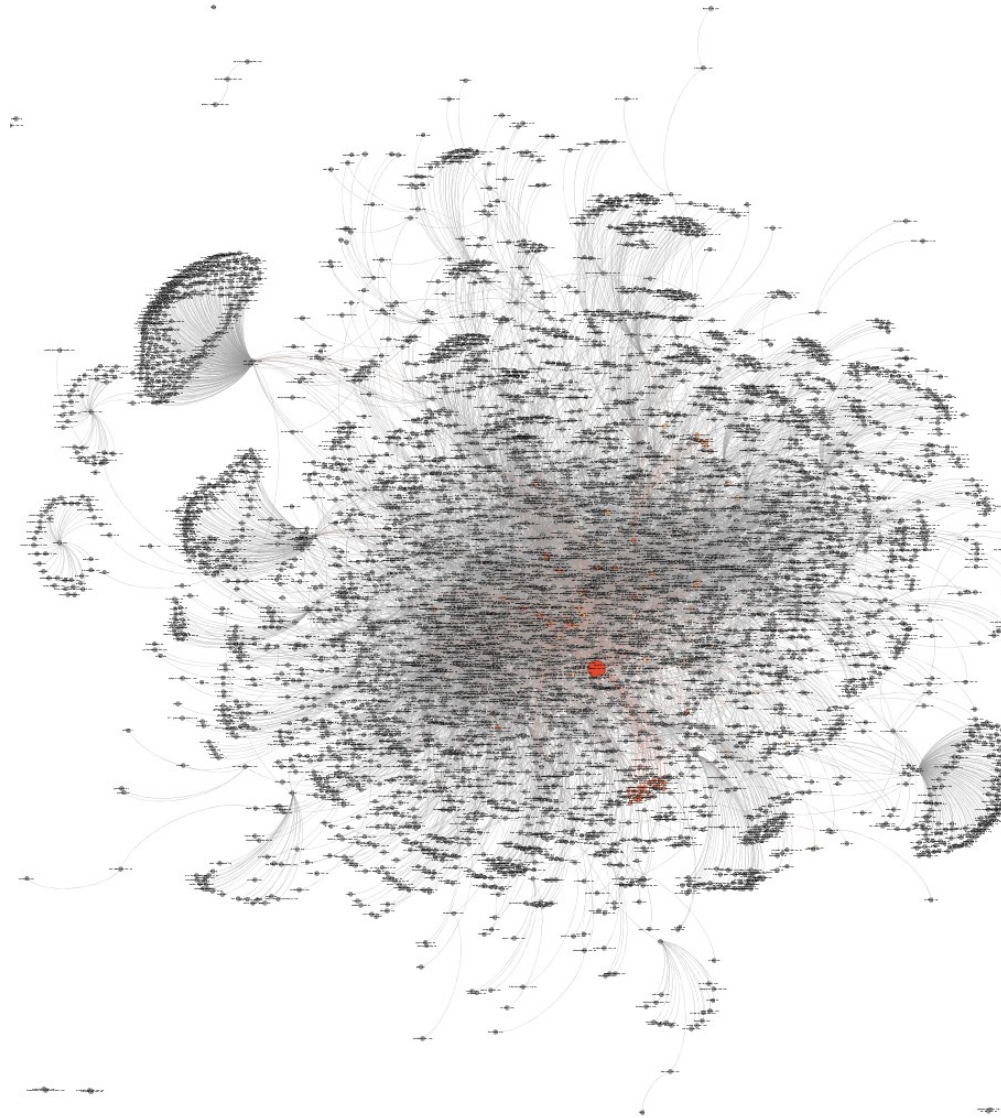
## Web as a directed graph:

- Nodes: Webpages
- Edges: Hyperlinks





## How to organise the web?



## How to organise the web?

### First try:

## Human Curated Web Directories Yahoo, DMOZ, LookSmart



What's New | Check Email | **YAHOO!** | My | ?  
 Personalize | Help

**Yahoo! Auctions**  
 bid & sell for free

**FAMILY GUY** Win 6 days in Hawaii!  
 Go to fox.com FOX

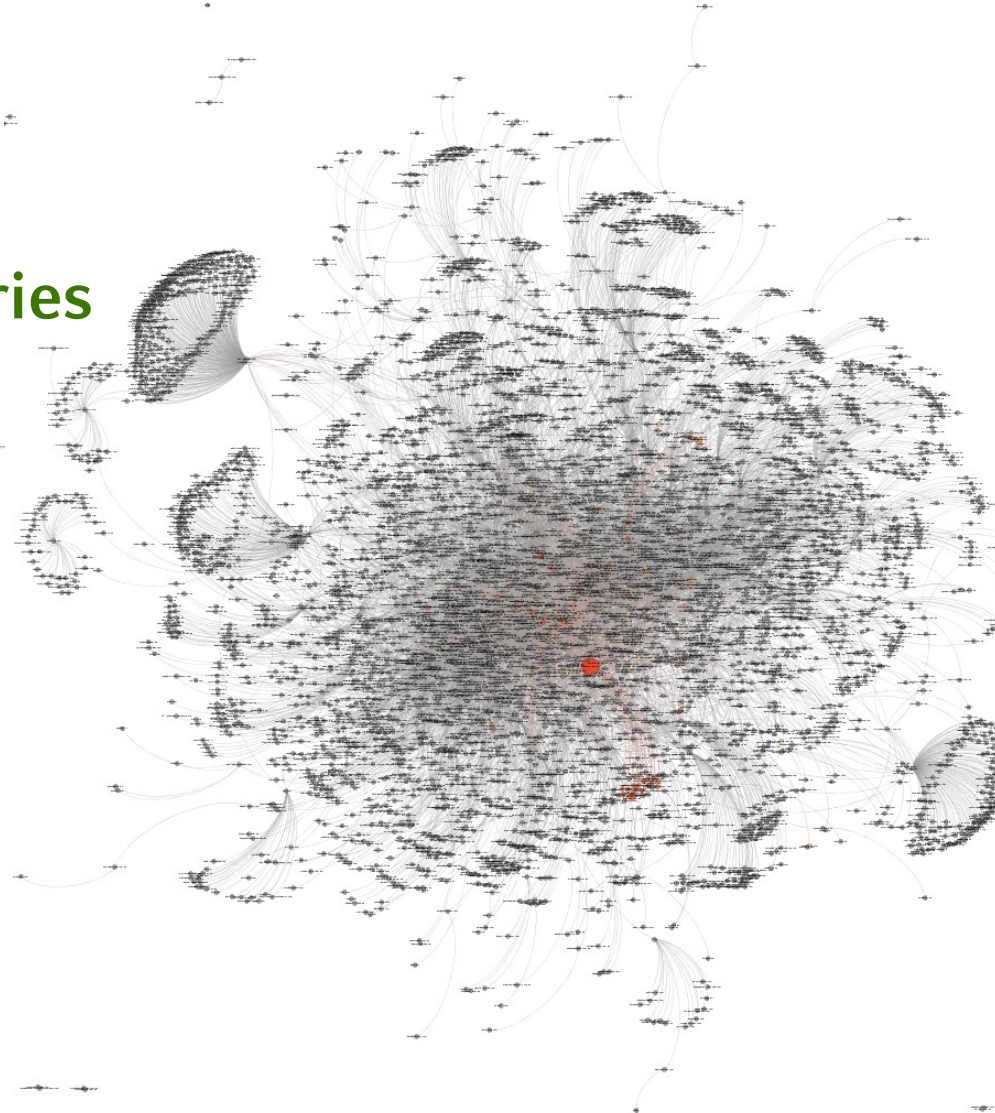
**Park Your Domain Free**

Search [advanced search](#)

**Yahoo! Mail** - Get your **free** e-mail account today!

[Shopping](#) - [Yellow Pages](#) - [People Search](#) - [Maps](#) - [Travel Agent](#) - [Classifieds](#) - [Personals](#) - [Games](#) - [Chat](#)  
[Email](#) - [Calendar](#) - [Pager](#) - [My Yahoo!](#) - [Today's News](#) - [Sports](#) - [Weather](#) - [TV](#) - [Stock Quotes](#) - [more...](#)

<b>Arts &amp; Humanities</b> <a href="#">Literature, Photography...</a>	<b>News &amp; Media</b> <a href="#">Full Coverage, Newspapers, TV...</a>	<b>In the News</b> <ul style="list-style-type: none"> <li>• <a href="#">King Hussein of Jordan dies</a></li> <li>• <a href="#">Online: Lewinsky video testimony</a></li> <li>• <a href="#">NASA comet mission</a></li> <li>• <a href="#">NBA season opens</a></li> <li>• <a href="#">Weekend's top movies</a></li> </ul> <a href="#">more...</a>
<b>Business &amp; Economy</b> <a href="#">Companies, Finance, Jobs...</a>	<b>Recreation &amp; Sports</b> <a href="#">Sports, Travel, Autos, Outdoors...</a>	<b>Inside Yahoo!</b> <ul style="list-style-type: none"> <li>• <a href="#">Y! Personals</a> - find a Valentine</li> <li>• <a href="#">Shop</a> for your Valentine</li> <li>• <a href="#">Y! Clubs</a> - create your own</li> </ul> <a href="#">more...</a>
<b>Computers &amp; Internet</b> <a href="#">Internet, WWW, Software, Games...</a>	<b>Reference</b> <a href="#">Libraries, Dictionaries, Quotations...</a>	
<b>Education</b> <a href="#">Universities, K-12, College Entrance...</a>	<b>Regional</b> <a href="#">Countries, Regions, US States...</a>	
<b>Entertainment</b> <a href="#">Cool Links, Movies, Humor, Music...</a>	<b>Science</b> <a href="#">Biology, Astronomy, Engineering...</a>	
<b>Government</b> <a href="#">Military, Politics, Law, Taxes...</a>	<b>Social Science</b> <a href="#">Archaeology, Economics, Languages...</a>	
<b>Health</b> <a href="#">Medicine, Diseases, Drugs, Fitness...</a>	<b>Society &amp; Culture</b> <a href="#">People, Environment, Religion...</a>	



**How to organise the web?**

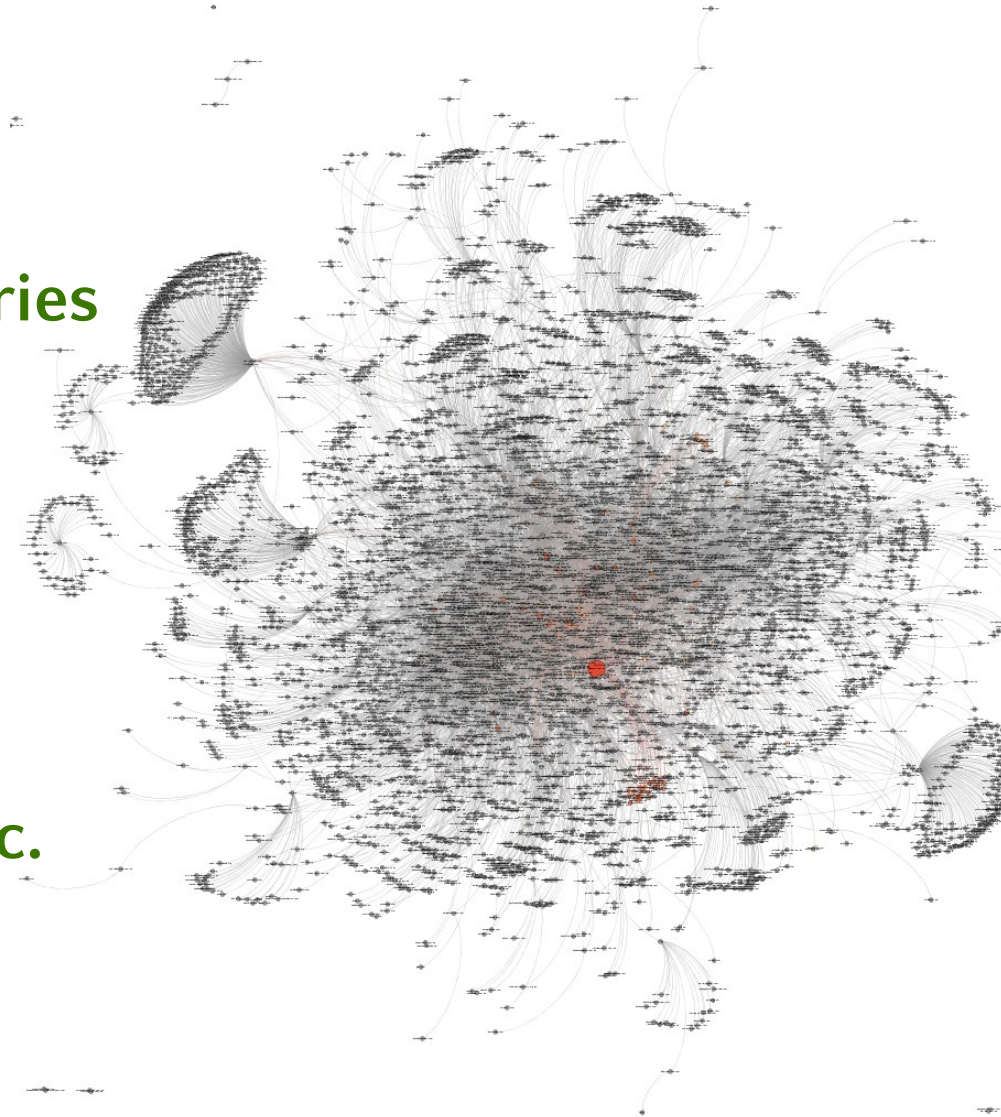
**First try:**

**Human Curated Web Directories**

**Second try:**

**Web Search**

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



- 1) Web contains many sources of information.  
→ Who to trust?

Idea: Trustworthy pages may point to each other

- 2) What is the “best” answer to a certain query?  
→ How to rank results?

No single right answer.

## Early Search Engines: Crawl the web, list terms, create inverted index

<http://www.example.org>

### Headline

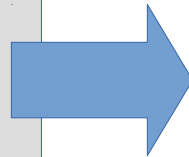
This text contains words.  
Words are important. Many  
words appear in this text.

## Early Search Engines: Crawl the web, list terms, create inverted index

http://www.example.org|

### Headline

This text contains words.  
Words are important. Many  
words appear in this text.



Problem:  
Term Spam

appear	example.org (1)
are	example.org (1)
contains	example.org (1)
headline	example.org (1)
important	example.org (1)
in	example.org (1)
many	example.org (1)
text	example.org (2)
this	example.org (2)
words	example.org (3)

# Web Search: Ranking Results

Not all web pages are equally "important"

[www.nytimes.com](http://www.nytimes.com)  
(The New York Times)

vs.

[www.thetimesonline.com](http://www.thetimesonline.com)  
(The Times of Northwest Indiana, Munster, IN)



The New York Times  
Tuesday, January 12, 2016

World U.S. Politics N.Y. Business Opinion Tech Science Health Sports Arts Style Food Travel Magazine T Magazine Real Estate ALL

**DEVELOPING**  
**Explosion Kills at Least 10 in Tourist District of Istanbul**  
By HELENA DUNN 9:04 AM ET  
President Recep Tayyip Erdogan said a Syrian suicide bomber was behind the attack in the Istanbul district of Sultanahmet that killed 10 and wounded at least 15 people, the latest in a series of terrorist assaults on Turkey.

**ELECTION 2016**  
Sen. Bernie Sanders of Vermont greets supporters at an event in Albany, N.Y., on Sunday. (AP/Wide World Photos)

**NEWS ANALYSIS**  
**What the President Probably Won't Say in Tonight's Address**  
By PETER SALES  
In his State of the Union speech, President Obama will offer reassurance on efforts against terrorism, but he is said to believe that the sense of danger has been inflated.

**MOVING**  
**Obama's Gun Policies Are Popular, Polls Show**  
By DAVID LEE  
A "Look Back at the First Lady's Guests" 9:13 AM ET

**FROM THE MAGAZINE**  
**The Trials of Alice Hoffman**  
By DEBORAH LEWIS-SEARS  
Her first book, "On the Run" — about the lives

**The Opinion Pages**  
OP-ED | RICHARD D. KAHLENBERG  
**Strong Unions, Strong Democracy**  
Weakening labor increases inequality and instability.  
• Brooks: The Brutalism of Ted Cruz  
• Rolfer: Ads That Play to Racist Fears  
• Editorial: DuPont Kept Using a Toxic Chemical  
• Room for Debate: How to Define Obama's Legacy  
• The Slope: David Bowie's Vision of Love

**Agony and Starvation in the Syrian War**  
By FRED BODENBERG  
Aid convoys and their supplies offer only a respite. What is needed is an end to the Syrian war.

**The Other Refugee Crisis**  
By BRUNO BRESSAN  
How do we support those who want to return to Syria?

**Watching**  
24m A medical panel has updated its mammogram guidelines but maintains that women with average risk can safely be going screening at 50, a stance that has inspired much debate. The New York Times  
3h China's Dalian Wanda Group said that it was buying **Legendary Entertainment**, one of Hollywood's highest movie production companies, for as much as \$3.5 billion. The New York Times  
5h The fantasy sports websites FanDuel and DraftKings will be allowed to operate in New York while they try to fend off a challenge from the state's attorney general. The New York Times  
5h Prosecutors in Florence, Italy, are investigating the death of an **American woman** who was found in her apartment over the weekend, the police said. The New York Times  
12h

**Four Teenagers Are Charged in Brooklyn Rape Case**



Customer Service E-notice Contact Us Calendar Join the conversation Log In Register Subscribe

nwi.com presented by THE TIMES  
Daily Forecast

News Sports Photos Business Opinion Entertainment Lifestyles Obit Health Cars Jobs Homes Deals Classifieds

Featured Updated Regional Crime Maps We want to celebrate your milestones Holiday wish lists for Region charities State, Region history GALLERIES: Faces of

**Interact: State of the Union quiz**  
ASSOCIATED PRESS  
Test your knowledge with a 10-question interactive quiz of State of the Union facts and facts.

2 hours ago  
QUIZ CRASH HEALTH WEATHER GG TRAVIA

**INDIANA HISTORICAL SOCIETY**  
Quiz: How well do you know Indiana history?  
Happy 20th birthday, Indiana! Take this quiz to see how well you know the history of the state. Who knows? Maybe you'll learn something new! ...  
January 12, 2016 10:00 am

**Latest Updates**  
CHARLES HAYNES: In 2016, genocide must end 4 minutes ago  
Sawyerman elected Crown Point council president 34 minutes ago  
Four new police officers hired for Cal City 1 hour ago  
Tri-Creek administrators want security feedback 2 hours ago

**Today's Deal**  
loading...  
**Most Popular**  
Articles Facebook

- South Haven woman accused of sexual relationship with boy
- UPDATE: icy roads, school delays in wake of overnight snow
- Charges filed in case of abused puppy
- UPDATE: Lansing Auto Zone robbed at gunpoint
- UPDATE: 1 to 3 inches of snow expected overnight

Not all web pages are equally “important”

[www.nytimes.com](http://www.nytimes.com)  
(The New York Times)

vs.

[www.thetimesonline.com](http://www.thetimesonline.com)  
(The Times of Northwest  
Indiana, Munster, IN)

in-links: ~13.600.000

in-links: 5.960

→ There is a large diversity in the web-graph node connectivity.

**IDEA: rank pages by their link structure!**



## Idea: links as votes

Page is more important if it has more links

**In-links? Out-links?**

## Idea: links as votes

Page is more important if it has more in-links

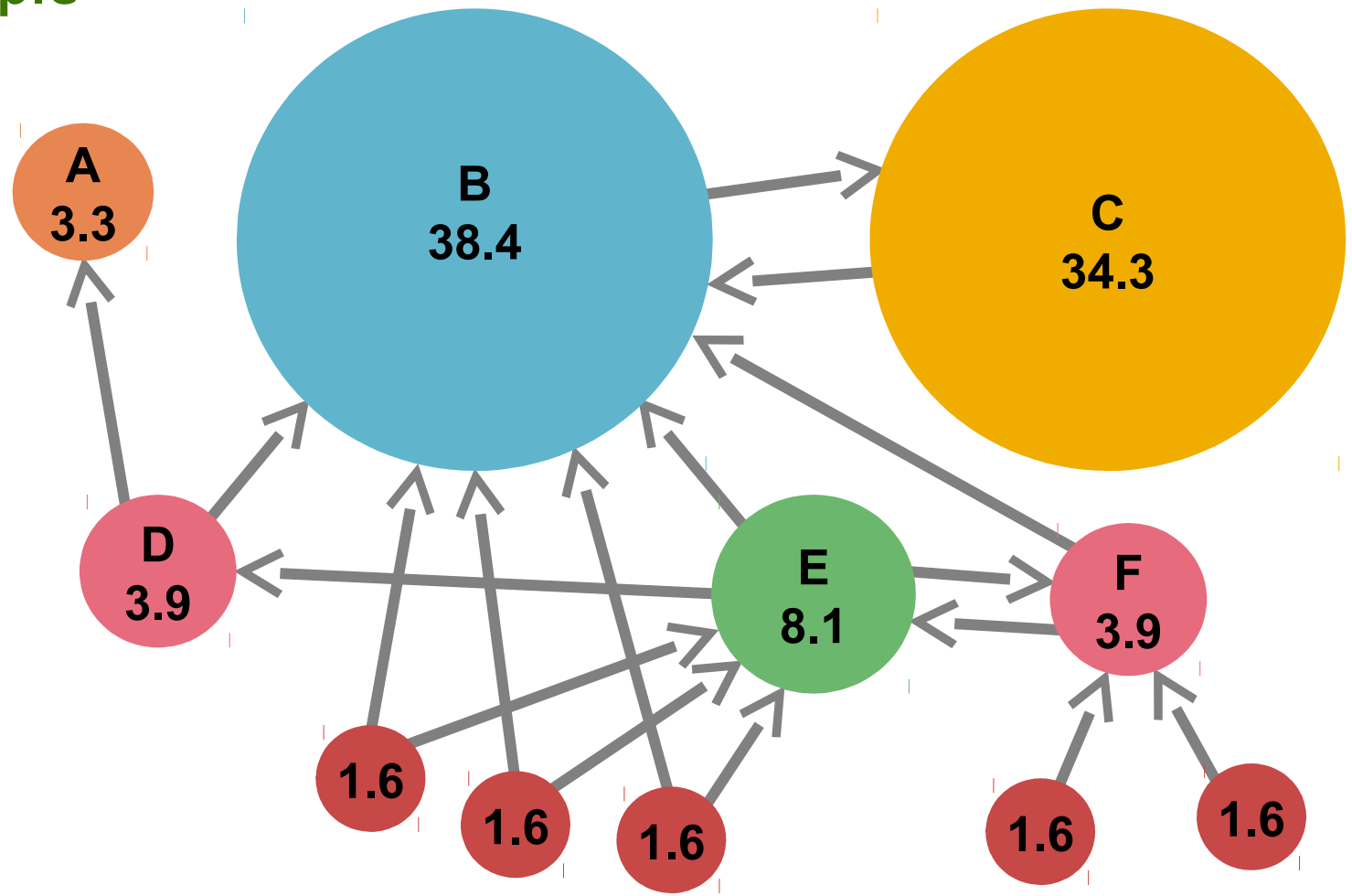
**Think of in-links as votes.**

## Are all in-links equal?

Links from important pages count more

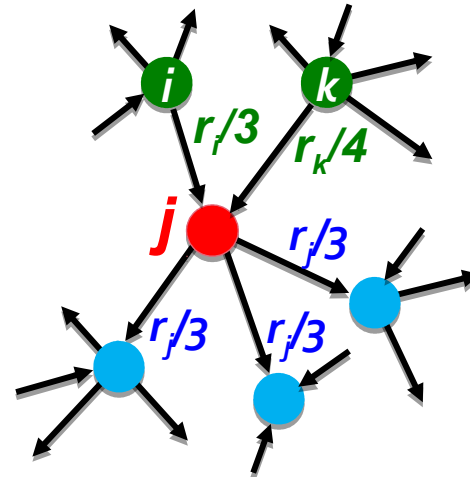
=> Recursive Definition!

## Example



- Each 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

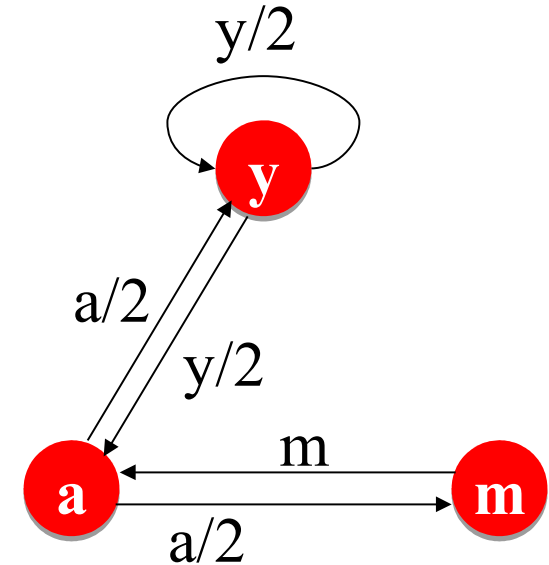
$$r_j = r_i/3 + r_k/4$$



- A "vote" from an important page is worth more
- A page is more important if it is pointed to by other important pages

Define a "rank"  $r_j$  for page  $j$   
(with  $d_i$  = out-degree of node  $i$ )

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



"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, no constants
  - No unique solution
  - All solutions equivalent modulo the scale factor
- Additional constraint forces uniqueness:
  - $r_y + r_a + r_m = 1$
  - Solution via Gaussian elimination  $r_y = 2/5, r_a = 2/5, r_m = 1/5$
- Gaussian elimination method works for small examples, but we need a better method for large web-sized graphs
- We need a new formulation!

- **Stochastic adjacency matrix  $M$** 
  - Let page  $i$  has  $d_i$  out-links
  - If  $i \rightarrow j$ , then  $M_{ji} = 1/d_i$ , else  $M_{ji} = 0$
  - $M$  is a column stochastic matrix: columns sum to 1
- **Rank vector  $r$ : vector with an entry per page**
  - $r_i$  is the importance score of page  $i$
  - $\sum_i r_i = 1$
- **The flow equations can be written**

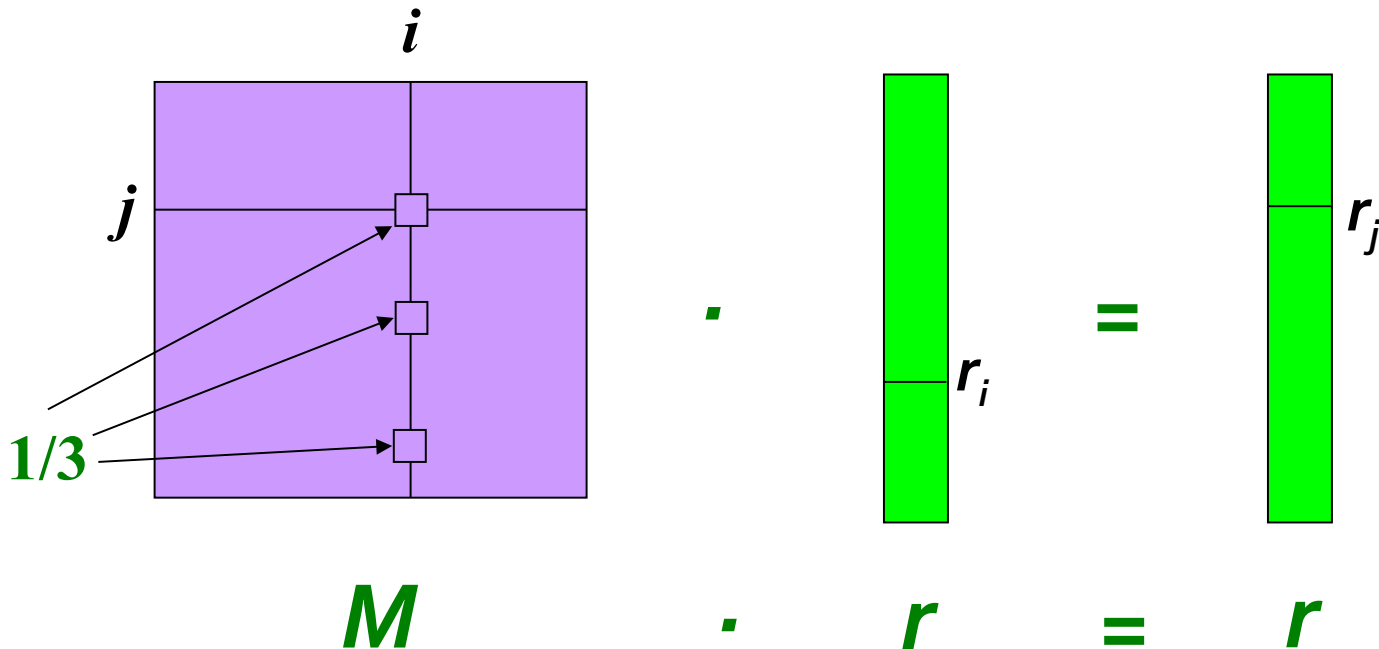
$$r = M \cdot r$$

# Example

● Remember the flow equation:  $r_j = \sum_{i \rightarrow j} \frac{r_i}{d_i}$

● Flow equation in matrix form:  $M \cdot r = r$

- Suppose page  $i$  links to 3 pages, including  $j$ :





- The flow equations can be written as

$$r = M \cdot r$$

- So the rank vector  $r$  is an *eigenvector* of the *stochastic web matrix*  $M$

- In fact, its first or principal *eigenvector* with corresponding *eigenvalue* 1
- Largest *eigenvalue* of  $M$  is 1 since  $M$  is column stochastic (with non-negative entries)
- We know  $r$  is unit length and each column of  $M$  sums to 1, so  $M \cdot r \leq 1$

Note:  
 $x$  is an eigenvector  
 with corresponding  
 eigenvalue  $\lambda$  if:

$$Ax = \lambda x$$

- We can now efficiently solve for  $r$ !  
**Power Iteration**

- **Power Iteration is an eigenvalue algorithm**
  - Also known as Von Mises iteration
  - Given a matrix  $A$ , P.I. returns a value  $\lambda$  and a nonzero vector  $v$ , such that  $Av = \lambda v$
- **Will find only the dominant eigenvector (the vector corresponding to the largest eigenvalue)**

$$\mathbf{r}^{(1)} = M \cdot \mathbf{r}^{(0)}$$

$$\mathbf{r}^{(2)} = M \cdot \mathbf{r}^{(1)} = M ( M \cdot \mathbf{r}^{(0)} ) = M^2 \cdot \mathbf{r}^{(0)}$$

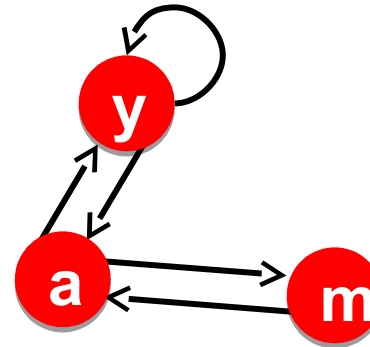
$$\mathbf{r}^{(3)} = M \cdot \mathbf{r}^{(2)} = M ( M^2 \cdot \mathbf{r}^{(0)} ) = M^3 \cdot \mathbf{r}^{(0)}$$

- Given a web graph with  $n$  nodes, where the nodes are pages and the edges are hyperlinks
- Power iteration: a simple iterative scheme
  - Suppose there are  $N$  web pages
  - Initialize:  $r^{(0)} = [1/N, \dots, 1/N]^T$
  - Iterate:  $r^{(t+1)} = M \cdot r^{(t)}$
  - Stop when:  $\| r^{(t+1)} - r^{(t)} \|_1 < \varepsilon$

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

## ● Power Iteration:

- Set  $r_j = 1/N$
- 1:  $r'_j = \sum_{i \rightarrow j} r_i / d_i$
- 2:  $r = r'$
- Goto 1



	y	a	m
y	1/2	1/2	0
a	1/2	0	1
m	0	1/2	0

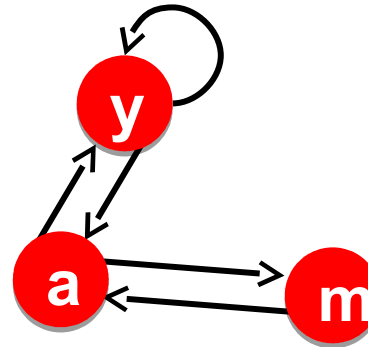
$$r_y = r_y / 2 + r_a / 2$$

$$r_a = r_y / 2 + r_m$$

$$r_m = r_a / 2$$

## ● Power Iteration:

- Set  $r_j = 1/N$
- 1:  $r'_j = \sum_{i \rightarrow j} r_i / d_i$
- 2:  $r = r'$
- Goto 1



	y	a	m
y	1/2	1/2	0
a	1/2	0	1
m	0	1/2	0

## ● Example:

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

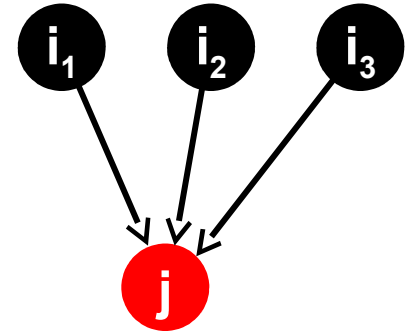
$$r_a = r_y/2 + r_m$$

$$r_m = r_a/2$$

$r_y$	=	1/3	1/3	5/12	9/24	6/15
$r_a$	=	1/3	3/6	1/3	11/24	... 6/15
$r_m$	=	1/3	1/6	3/12	1/6	3/15

## ● 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 page  $j$  linked from  $i$
- Process repeats indefinitely



$$r_j = \sum_{i \rightarrow j} \frac{r_i}{d_{\text{out}}(i)}$$

## ● Let:

- $p(t)$  ... vector whose  $i^{\text{th}}$  coordinate is the probability that surfer is at page  $i$  at time  $t$
- So,  $p(t)$  is a probability distribution over pages

- Where is surfer at time  $t + 1$ ?

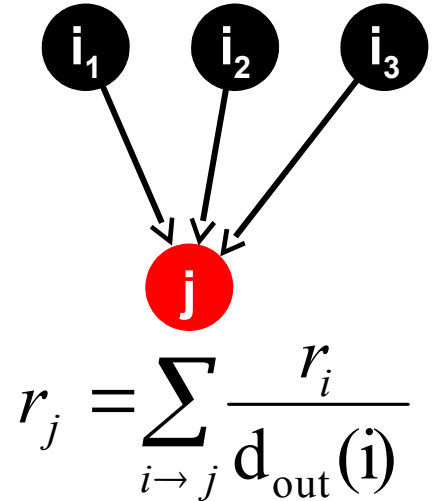
- Follows a link uniformly at random
- $$p(t + 1) = M \cdot p(t)$$

- 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 a random walk



A central result from the theory of random walks  
(a.k.a. Markov processes):

For graphs that satisfy **certain conditions**,  
the **stationary distribution is unique** and eventually  
will be reached no matter what the initial probability  
distribution at time  $t = 0$ .

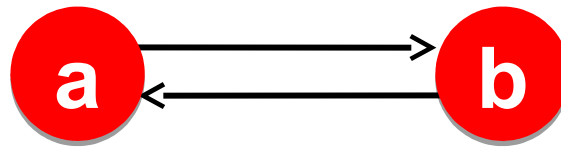


$$r_j^{(t+1)} = \sum_{i \rightarrow j} \frac{r_i^{(t)}}{d_i}$$

$$r = Mr$$

- Does this converge?
- Does it converge to what we want?
- Are results reasonable?

# Does this converge?

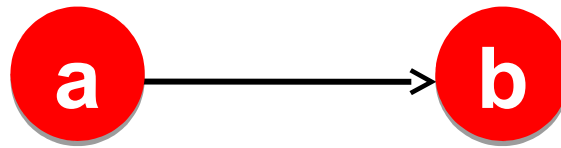


$$r_j^{(t+1)} = \sum_{i \rightarrow j} \frac{r_i^{(t)}}{d_i}$$

## ● Example:

$r_a$	1	0	1	0	1	0	1	0	...
$r_b$	0	1	0	1	0	1	0	1	

# Does it converge to what we want?



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

## ● Example:

$r_a$	1	0	0	0	0	0	0	...
$r_b$	0	1	0	0	0	0	0	

## 2 Problems:

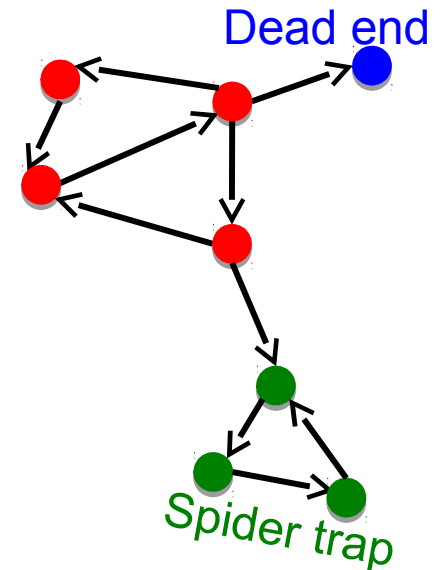
- **Some pages are dead ends  
(have no out-links)**

- Random walk has "nowhere to go" to
- Such pages cause "leak" of importance

- **Spider traps**

- (all out-links are within a group)

- Random walk gets "stuck" in a trap
- Eventually spider trap absorbs all importance

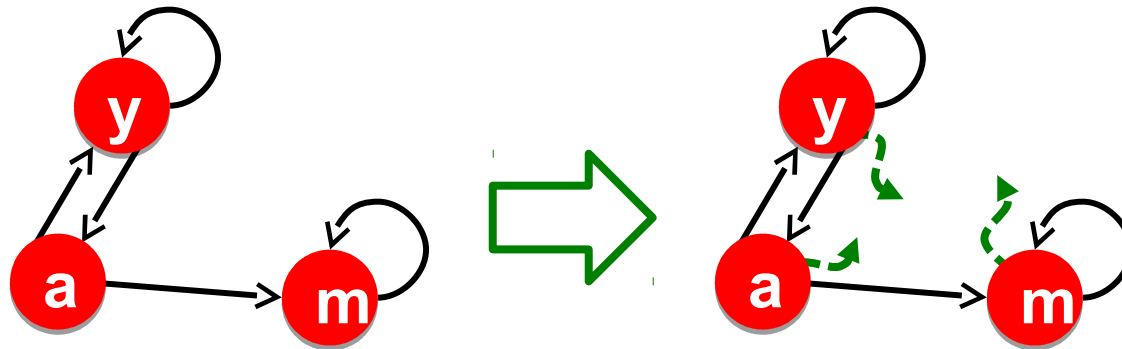


## The Google solution for spider traps: *Teleports*

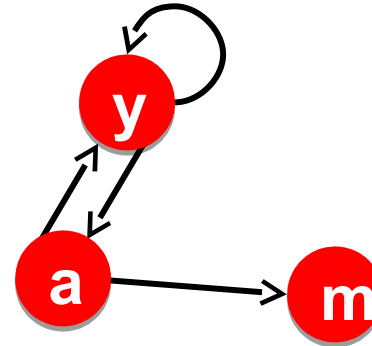
At each time step, the random surfer has two options:

- With probability  $\beta$ , follow a link at random
- With probability  $1 - \beta$ , jump to some random page
- Common values for  $\beta$  range between 0.8 and 0.9

Surfer will teleport out of spider trap within a few time steps



Dead ends cause the page importance to leak out, because the adjacency matrix is non-stochastic.



	y	a	m
y	1/2	1/2	0
a	1/2	0	0
m	0	1/2	0

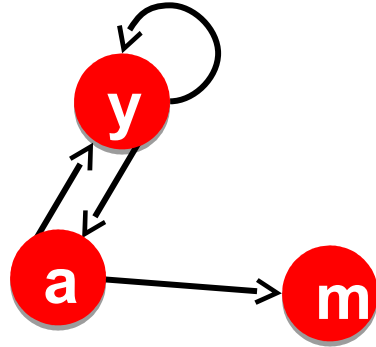
$$r_y = r_y / 2 + r_a / 2$$

$$r_a = r_y / 2$$

$$r_m = r_a / 2$$

# Dead Ends: Solution

Dead ends cause the page importance to leak out, because the adjacency matrix is non-stochastic.



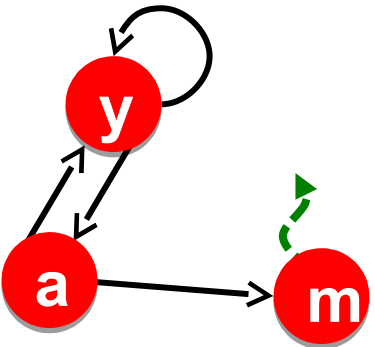
	y	a	m
y	1/2	1/2	0
a	1/2	0	0
m	0	1/2	0

Solution: Always teleport!  
Adjust matrix accordingly:

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

$$r_a = r_y / 2$$

$$r_m = r_a / 2$$



	y	a	m
y	1/2	1/2	1/3
a	1/2	0	1/3
m	0	1/2	1/3

## The final version of the Google PageRank: *[Brin-Page 98]*

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

(This formulation assumes  $M$  has no dead ends.  
 $M$  can either be preprocessed to remove all dead ends  
 or with explicit teleports to random links from dead ends.)

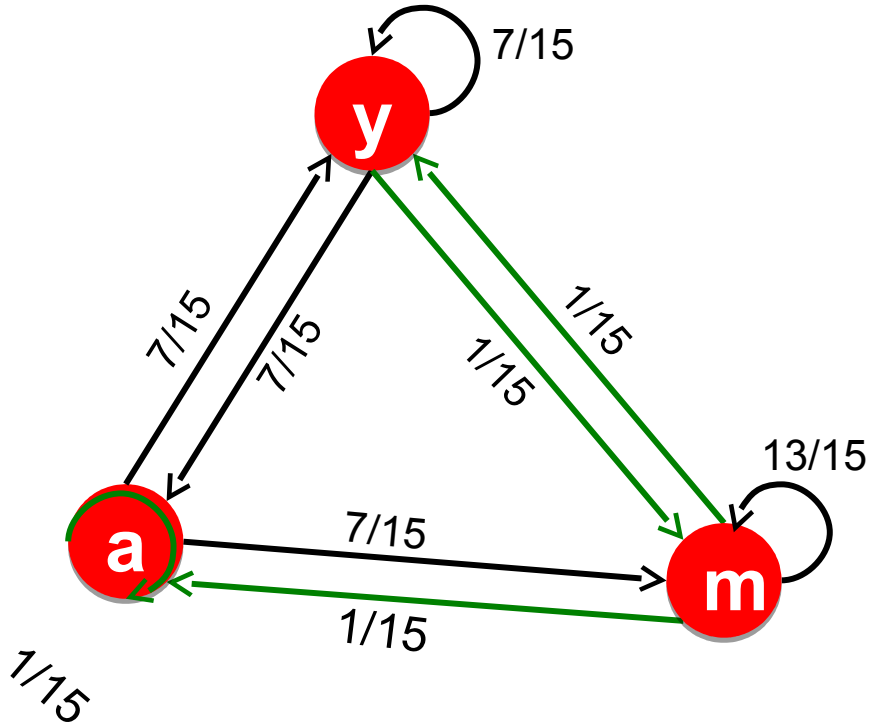


Google matrix  $A$  combines the adjacency matrix  $M$  with the random teleports by a factor  $\beta$ .

(With  $\beta = 0.8$  for this example)

$$\beta \begin{matrix} & \mathbf{M} \\ \begin{matrix} y \\ a \\ m \end{matrix} & \begin{bmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 0 & 0 \\ 0 & 1/2 & 1 \end{bmatrix} \end{matrix} + (1 - \beta) \begin{matrix} & \mathbf{[1/N]_{N \times N}} \\ \begin{matrix} y \\ a \\ m \end{matrix} & \begin{bmatrix} 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \end{bmatrix} \end{matrix} \\
 \begin{matrix} y \\ a \\ m \end{matrix} \begin{bmatrix} 7/15 & 7/15 & 1/15 \\ 7/15 & 1/15 & 1/15 \\ 1/15 & 7/15 & 13/15 \end{bmatrix} \\
 \mathbf{A}$$

# The Google Matrix



$$0.8 \begin{bmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 0 & 0 \\ 0 & 1/2 & 1 \end{bmatrix} + 0.2 \begin{bmatrix} 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \end{bmatrix}$$

$$\begin{matrix} y \\ a \\ m \end{matrix} \begin{bmatrix} 7/15 & 7/15 & 1/15 \\ 7/15 & 1/15 & 1/15 \\ 1/15 & 7/15 & 13/15 \end{bmatrix}$$

**A**

y	=	1/3	0.33	0.24	0.26		7/33
a		1/3	0.20	0.20	0.18	...	5/33
m		1/3	0.46	0.52	0.56		21/33

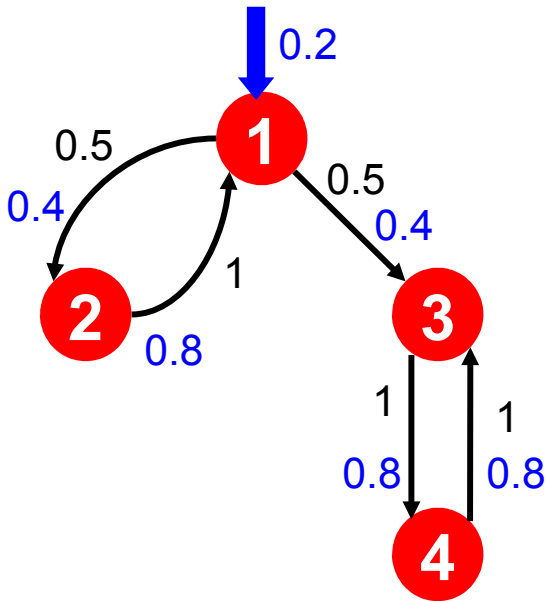
# Some Problems with PageRank

- **Measures generic popularity of a page**
  - Biased against topic-specific authorities
  - Solution: Topic-specific PageRank
- **Uses only one measure of importance**
  - Other models exist
  - Solution: e.g., Hubs and Authorities
- **Susceptible to Link Spam**
  - Evolved from term spam (see: older search engines)
  - Artificial link topographies created to boost page rank
  - Solution: TrustRank

- Instead of generic popularity, can we measure popularity within a certain topic?
- Goal: evaluate web pages not only according to their popularity, but by how close they are to a particular topic, e.g., “sports” or “history”
- Allows search queries to be answered based on user interest
  - Example: Query “Trojan” may yield different results depending on whether user is interested in sports, history, computer security, ...

- **Modification in random walk behaviour (teleports)**
- **Teleport has probability to go to:**
  - Standard PageRank: Any page with equal probability  
To avoid dead ends and spider-traps
  - Topic-specific PageRank: A topic specific set of “relevant” pages (teleport set)
- **Idea: Bias the random walk**
  - When walker teleport, they pick a page from set  $S$
  - $S$  contains only pages that are relevant to the topic, e.g., from Open Directory (DMOZ) pages for given topic
  - For each teleport set  $S$ , we get a different vector  $r_s$

# Example: Topic-specific PageRank



Suppose  $S = \{1\}$ ,  $\beta = 0.8$

Node	Iteration				
	0	1	2	...	stable
1	0.25	0.4	0.28		0.294
2	0.25	0.1	0.16		0.118
3	0.25	0.3	0.32		0.327
4	0.25	0.2	0.24		0.261

$S = \{1\}$ ,  $\beta = 0.90$ :

$r = [0.17, 0.07, 0.40, 0.36]$

$S = \{1\}$ ,  $\beta = 0.8$ :

$r = [0.29, 0.11, 0.32, 0.26]$

$S = \{1\}$ ,  $\beta = 0.70$ :

$r = [0.39, 0.14, 0.27, 0.19]$

$S = \{1, 2, 3, 4\}$ ,  $\beta = 0.8$ :

$r = [0.13, 0.10, 0.39, 0.36]$

$S = \{1, 2, 3\}$ ,  $\beta = 0.8$ :

$r = [0.17, 0.13, 0.38, 0.30]$

$S = \{1, 2\}$ ,  $\beta = 0.8$ :

$r = [0.26, 0.20, 0.29, 0.23]$

$S = \{1\}$ ,  $\beta = 0.8$ :

$r = [0.29, 0.11, 0.32, 0.26]$

- **Create different PageRanks for different topics**
  - The 16 DMOZ top-level categories  
art, business, sports, ...
  
- **Which topic ranking to use?**
  - User can pick from a menu
  - Classify query into a topic
  - Use context of query:  
e.g., query is launched from website about  
certain topic, or history of queries
  - User context, e.g., bookmarks, ...

## ● “Normal” PageRank

- Teleports uniformly at random to any node
- All nodes have the same landing probability

$$S = [ 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1 ]$$

## ● Topic-specific PageRank, also known as Personalized PageRank

- Teleports to a topic specific set of pages
- Nodes can have different landing probabilities

$$S = [ 0.1, 0.0, 0.2, 0.0, 0.0, 0.0, 0.5, 0.0, 0.2, 0.0 ]$$

## ● Random walk with restarts

- Topic-specific with teleports to always the same node

$$S = [ 0.0, 0.0, 0.0, \mathbf{1.0}, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0 ]$$



- **Spamming:**  
Any deliberate action with the intent to boost a web page's position in search engine results incommensurate with page's actual relevance
- **Spam:**  
Query results that are the result of spamming  
→ very broad definition
- **Approximately 10% – 15% of web pages are spam**

- **Early spamming techniques flooded web pages with unfitting words to exploit search engines**
  - Example: Web page for T-Shirts includes the word “movie” over and over again
  - “Term spam”
- **As Google became more dominant, spam farms tried to target PageRank to a single page by placing many contextual links on other pages**
  - “Link Spam” or “Google Bomb”



**Web** Results 1 - 10 of about 969,000 for miserable failure. (0.06 seconds)

## Biography of President George W. Bush

Biography of the president from the official White House web site.

[www.whitehouse.gov/president/gwbbio.html](http://www.whitehouse.gov/president/gwbbio.html) - 29k - [Cached](#) - [Similar pages](#)

[Past Presidents](#) - [Kids Only](#) - [Current News](#) - [President](#)

[More results from www.whitehouse.gov »](#)

## Welcome to MichaelMoore.com!

Official site of the gadfly of corporations, creator of the film Roger and Me and the television show The Awful Truth. Includes mailing list, message board, ...

[www.michaelmoore.com/](http://www.michaelmoore.com/) - 35k - Sep 1, 2005 - [Cached](#) - [Similar pages](#)

## BBC NEWS | Americas | 'Miserable failure' links to Bush

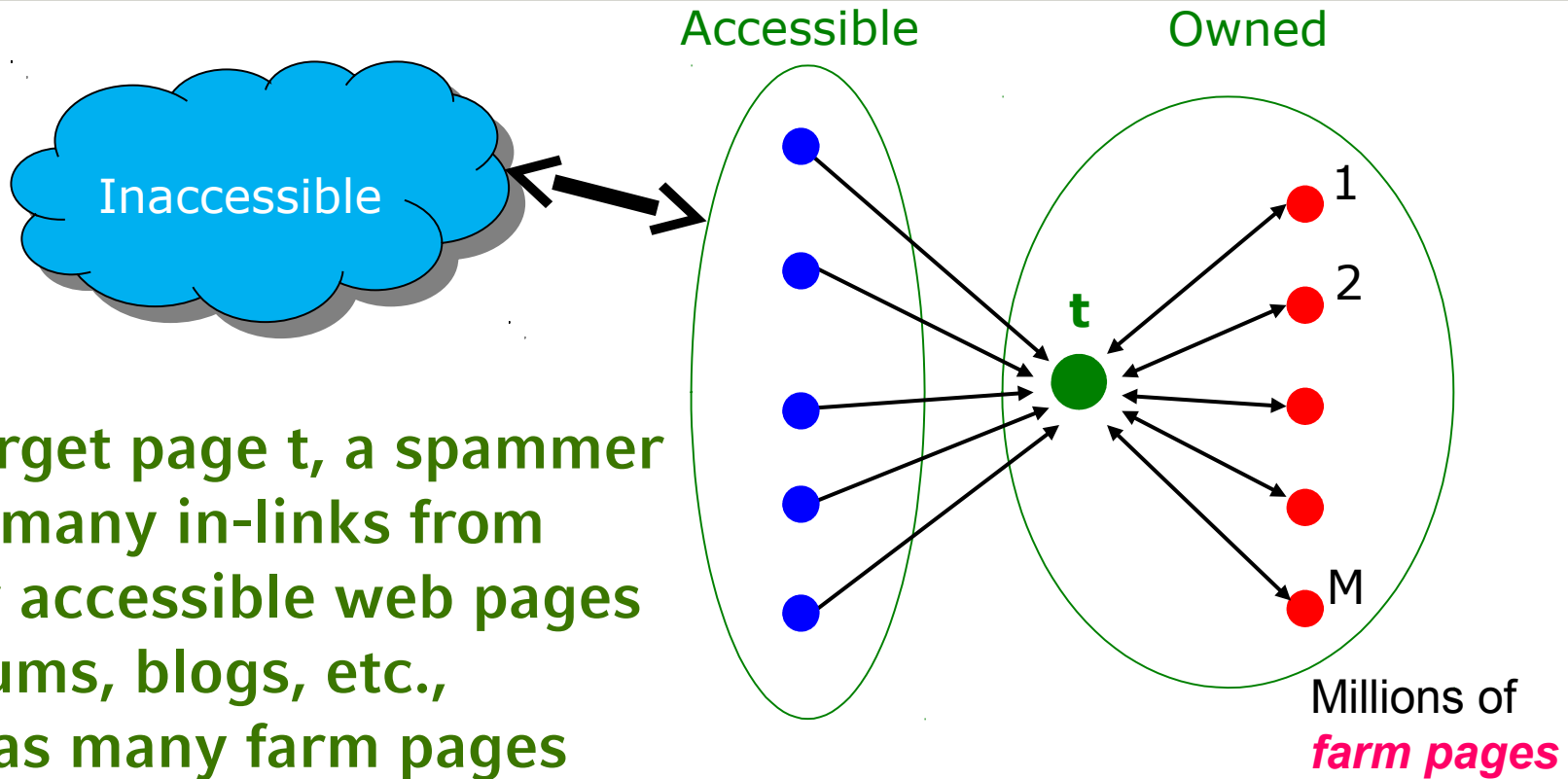
Web users manipulate a popular search engine so an unflattering description leads to the president's page.

[news.bbc.co.uk/2/hi/americas/3298443.stm](http://news.bbc.co.uk/2/hi/americas/3298443.stm) - 31k - [Cached](#) - [Similar pages](#)

## Google's (and Inktomi's) Miserable Failure

A search for **miserable failure** on Google brings up the official George W. Bush biography from the US White House web site. Dismissed by Google as not a ...

[searchenginewatch.com/sereport/article.php/3296101](http://searchenginewatch.com/sereport/article.php/3296101) - 45k - Sep 1, 2005 - [Cached](#) - [Similar pages](#)



For a target page  $t$ , a spammer creates many in-links from publicly accessible web pages like forums, blogs, etc., as well as many farm pages on own infrastructure to create a closely connected clique.

## ● Combating Term Spam:

- Analyze text using statistical methods
- Similar to email spam filtering
- Detecting duplicate pages

## ● Combating Link Spam:

- Detection and blacklisting of structures that look like spam farms
- Leads to another war: hiding and detecting
- TrustRank = topic-specific PageRank with teleport to a set of trusted pages, e.g., .edu domains or similar

## ● Alternative model for TrustRank: Trust Propagation

Initial seed set of trusted pages (evaluated by hand)

### ● Set trust $tp$ of each trusted page $p$ to 1

- For each out-link from  $p$ , a portion of the trust is passed on to target page  $q$

### ● Trust is additive

- Trust of  $q$  is sum of all trust conferred by its in-links

### ● If trust is below a threshold, page is flagged as spam