

Link Analysis Algorithms

Page Rank

A. Farzindar
INF 553

With slide contributions from
J. Leskovec, Anand Rajaraman, Jeffrey D. Ullman, Wensheng Wu

New Topic: Graph Data

High dim. data

Locality sensitive hashing

Clustering

Dimensionality reduction

Graph data

PageRank, SimRank

Community Detection

Spam Detection

Infinite data

Filtering data streams

Web advertising

Queries on streams

Machine learning

SVM

Decision Trees

Perceptron, kNN

Apps

Recommender systems

Association Rules

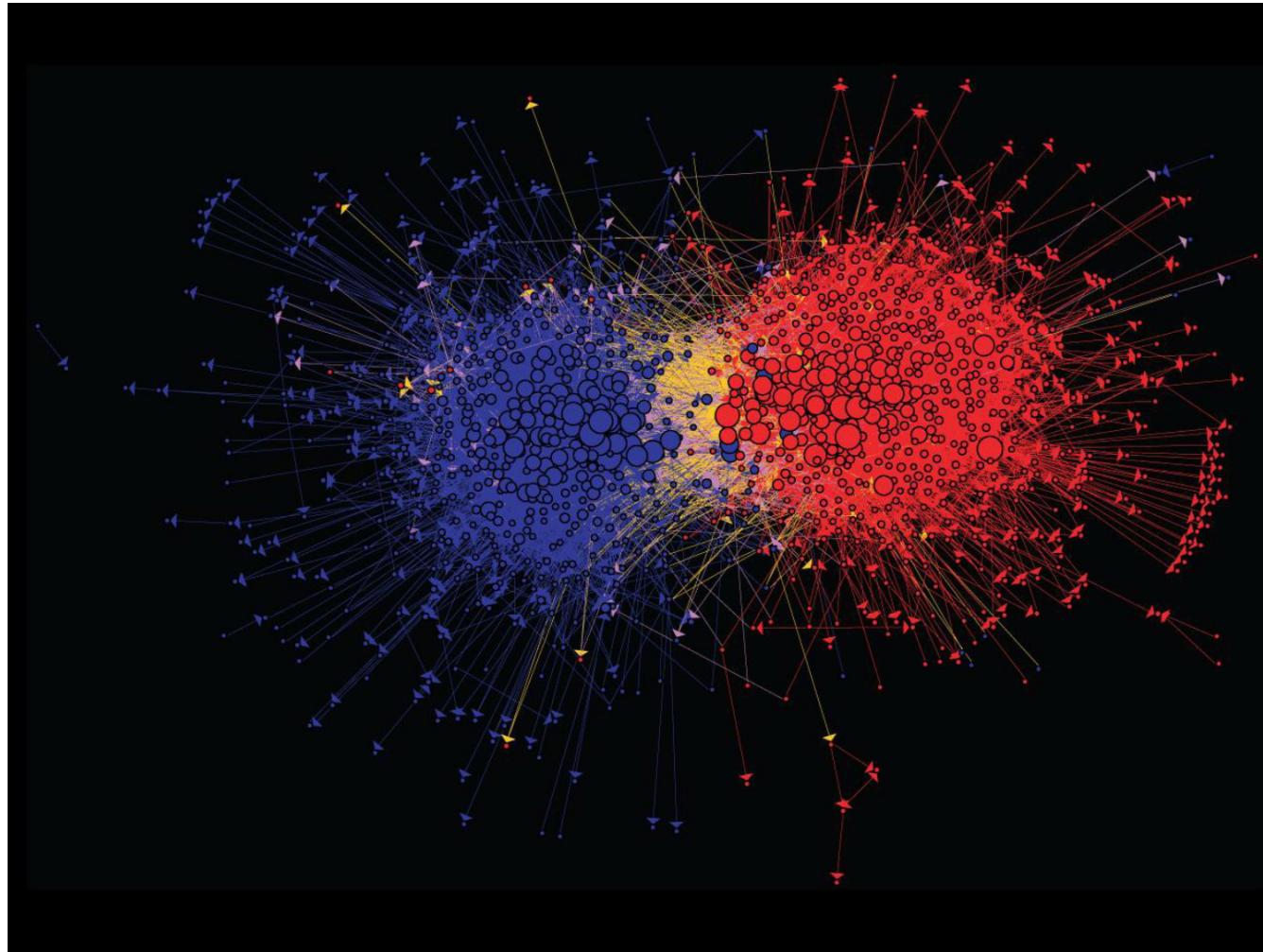
Duplicate document detection

Graph Data: Social Networks



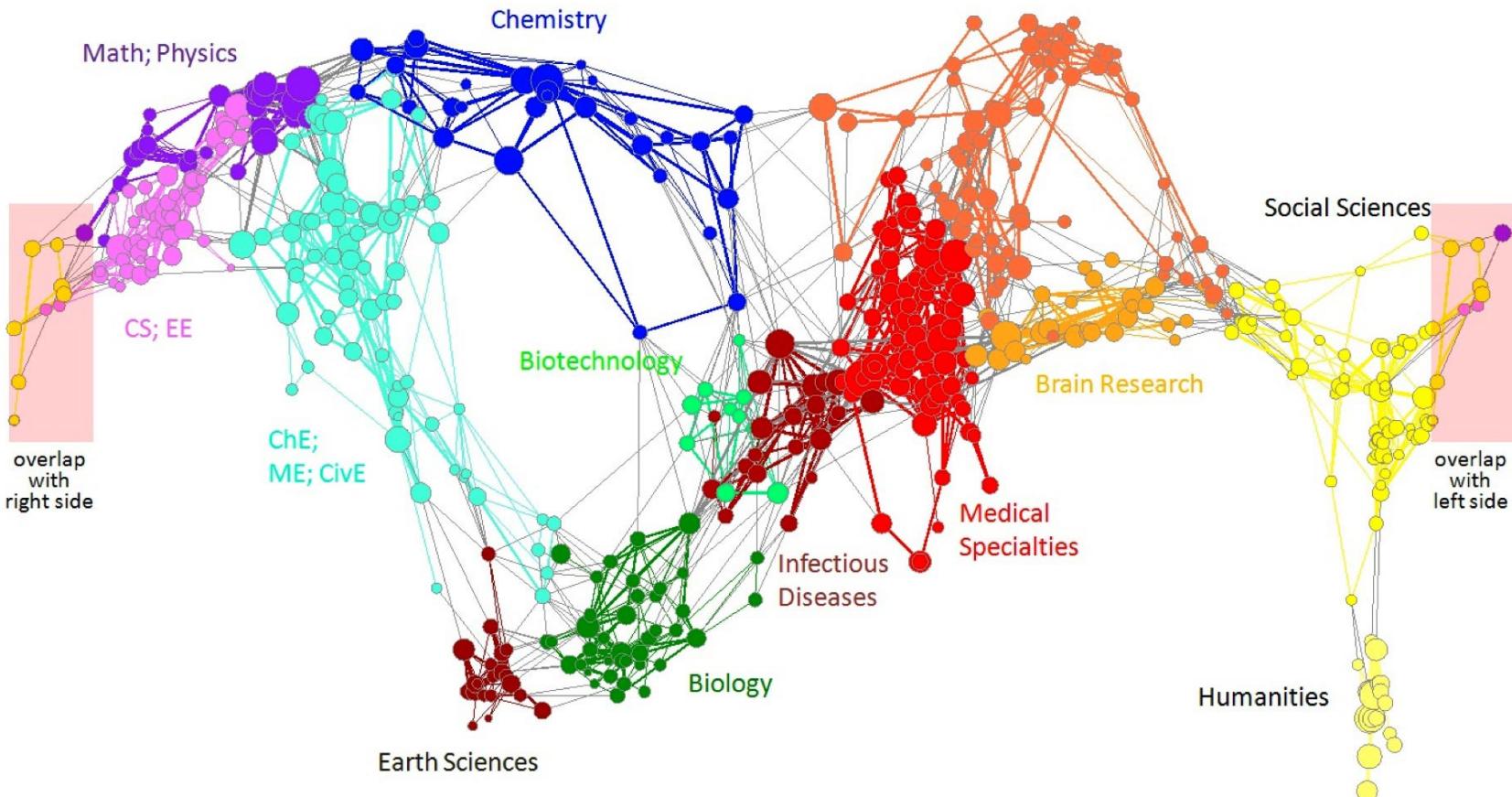
Facebook social graph
4-degrees of separation [Backstrom-Boldi-Rosa-Ugander-Vigna, 2011]

Graph Data: Media Networks



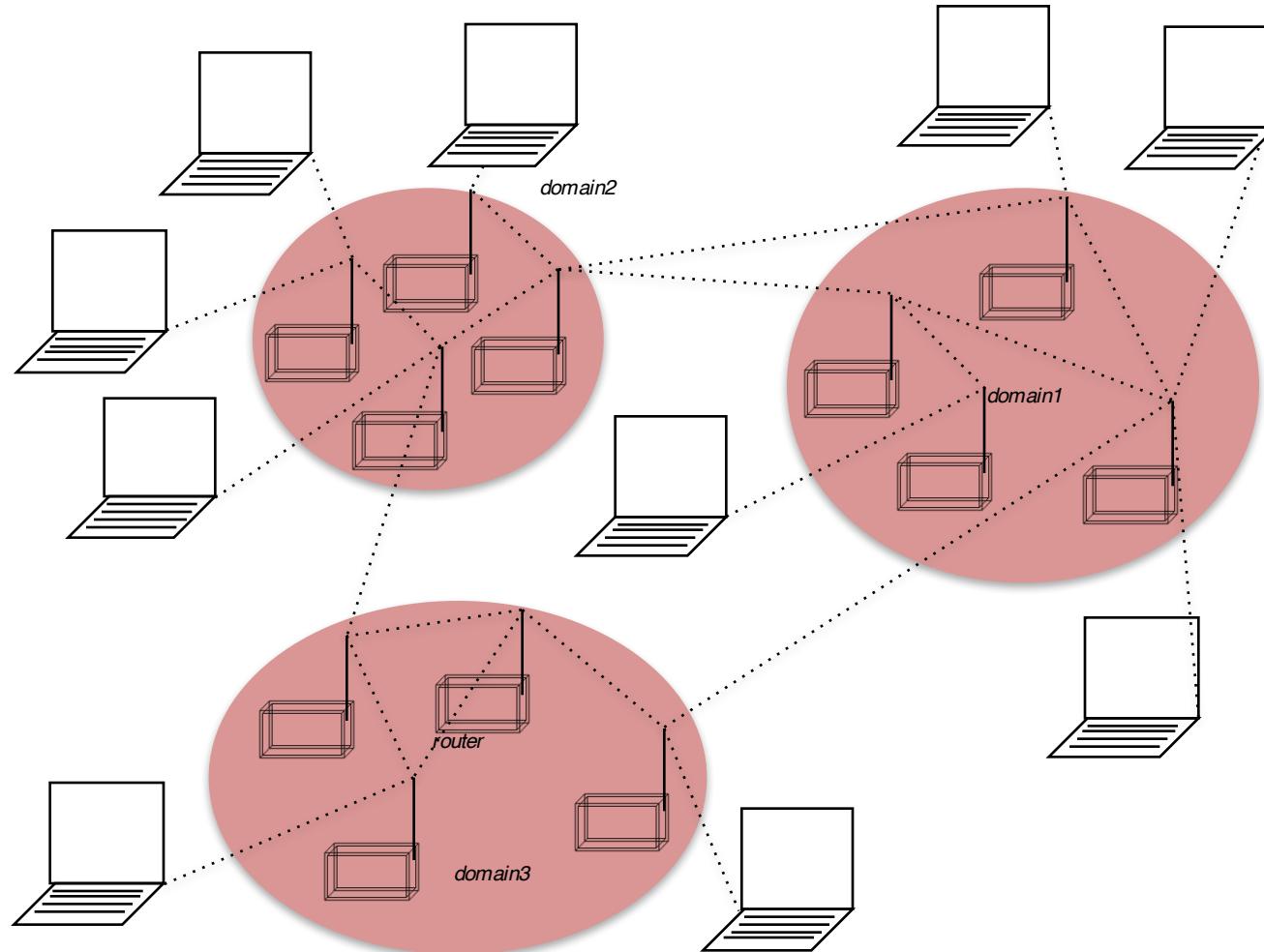
Connections between political blogs
Polarization of the network [Adamic-Glance, 2005]⁴

Graph Data: Information Nets



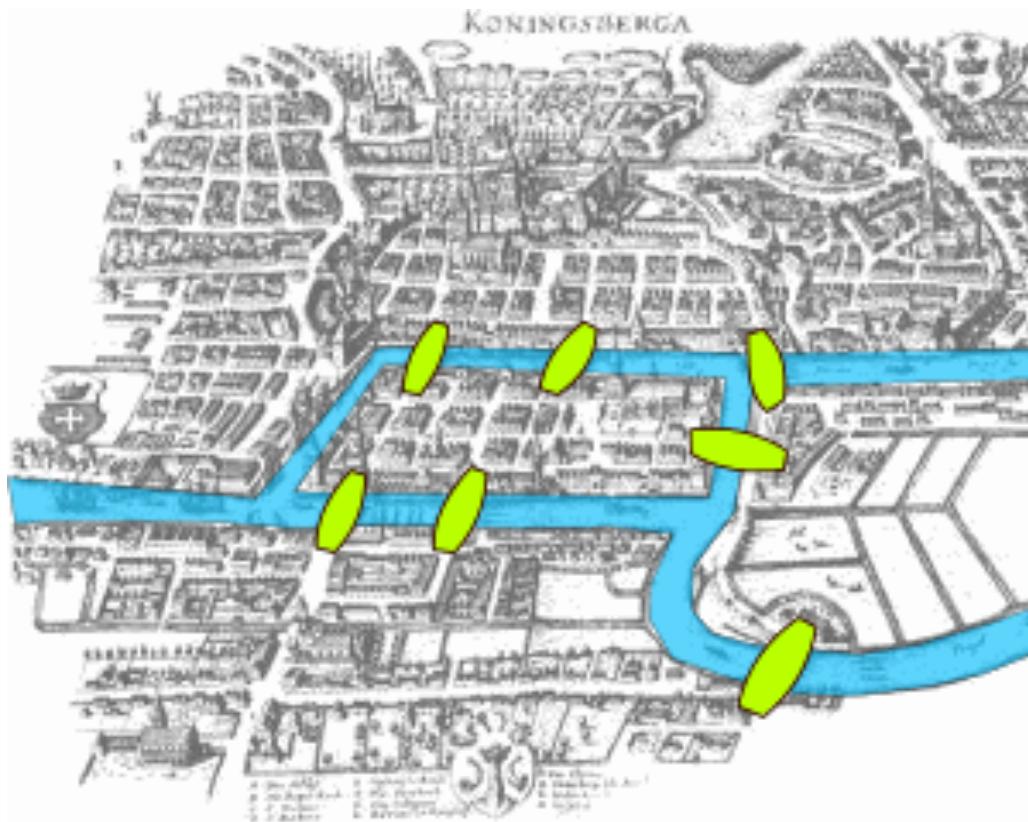
Citation networks and Maps of science
[Börner et al., 2012]

Graph Data: Communication Nets



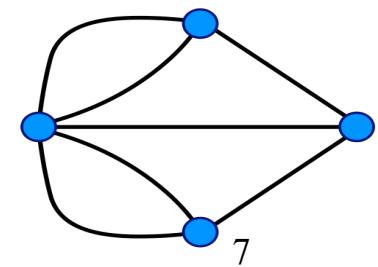
Internet

Graph Data: Technological Networks



**Seven Bridges of
Königsberg
[Euler, 1735]**

Return to the starting point by traveling each link of the graph once and only once.



The Problem

Web as a Graph

◆ Web as a directed graph:

- Nodes: Webpages
- Edges: Hyperlinks

I teach a
class on
Networks.

CS555:
Classes are
in the
SAL
building

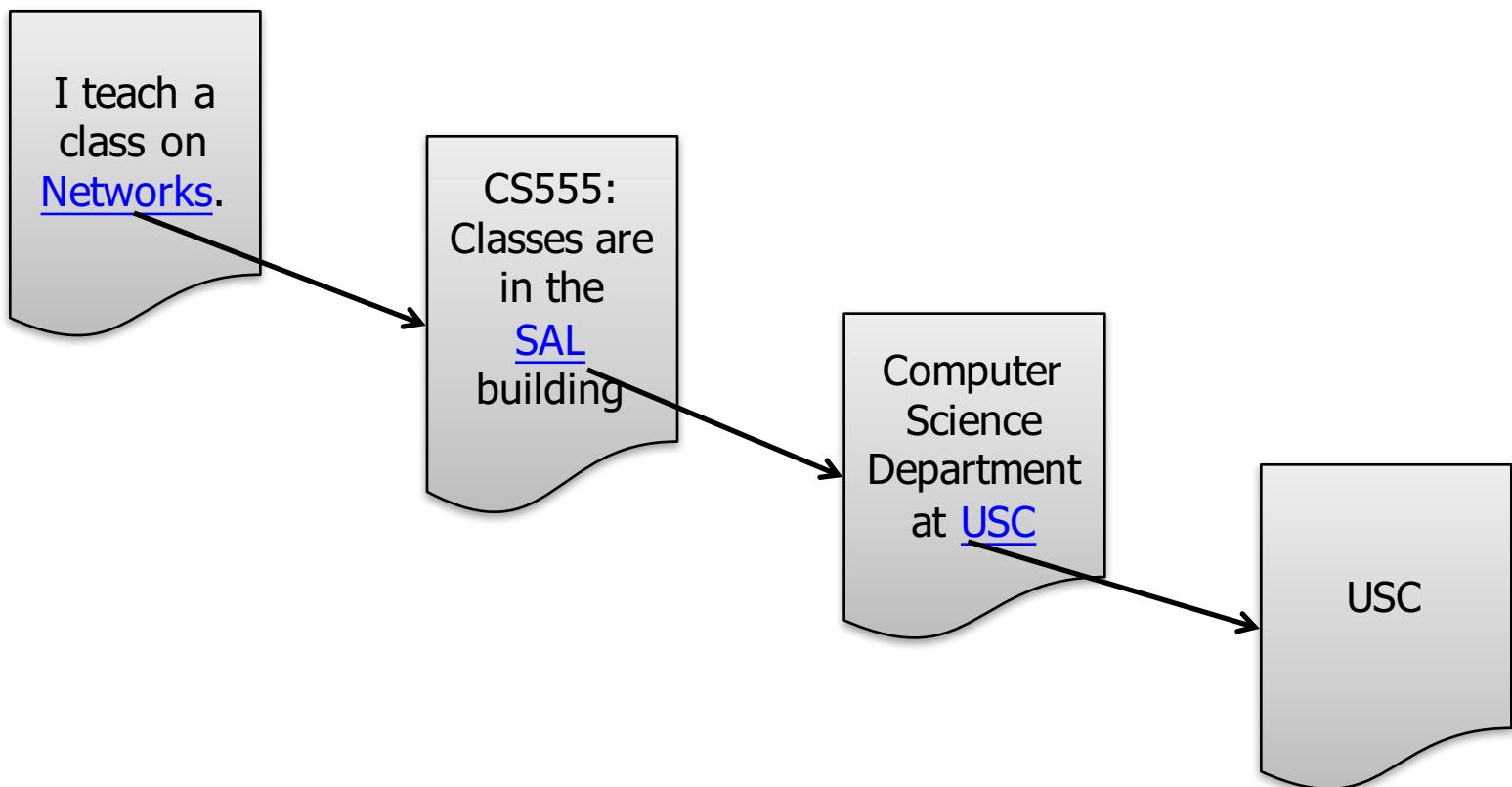
Computer
Science
Department
at USC

USC

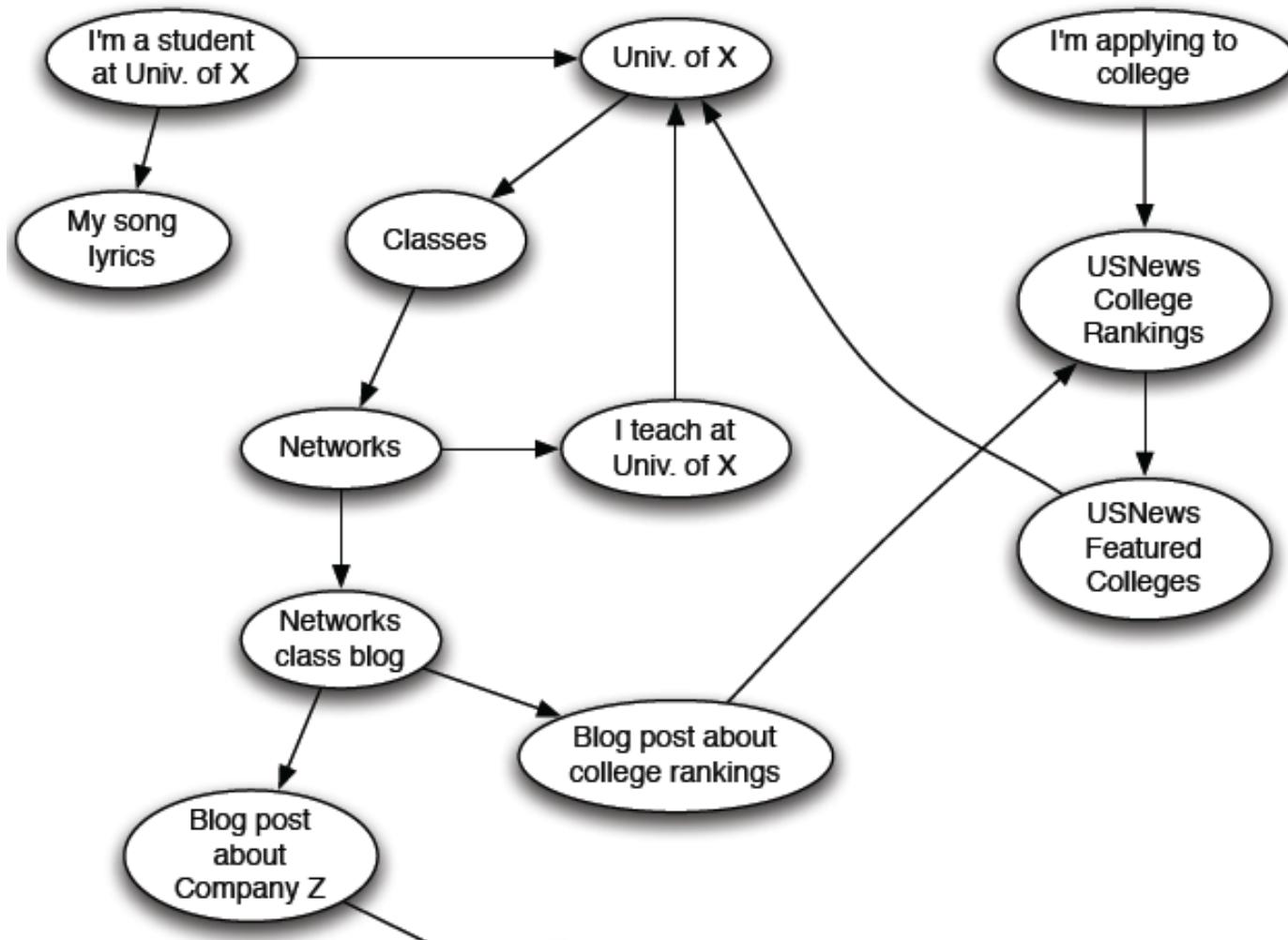
Web as a Graph

◆ Web as a directed graph:

- Nodes: Webpages
- Edges: Hyperlinks



Web as a Directed Graph

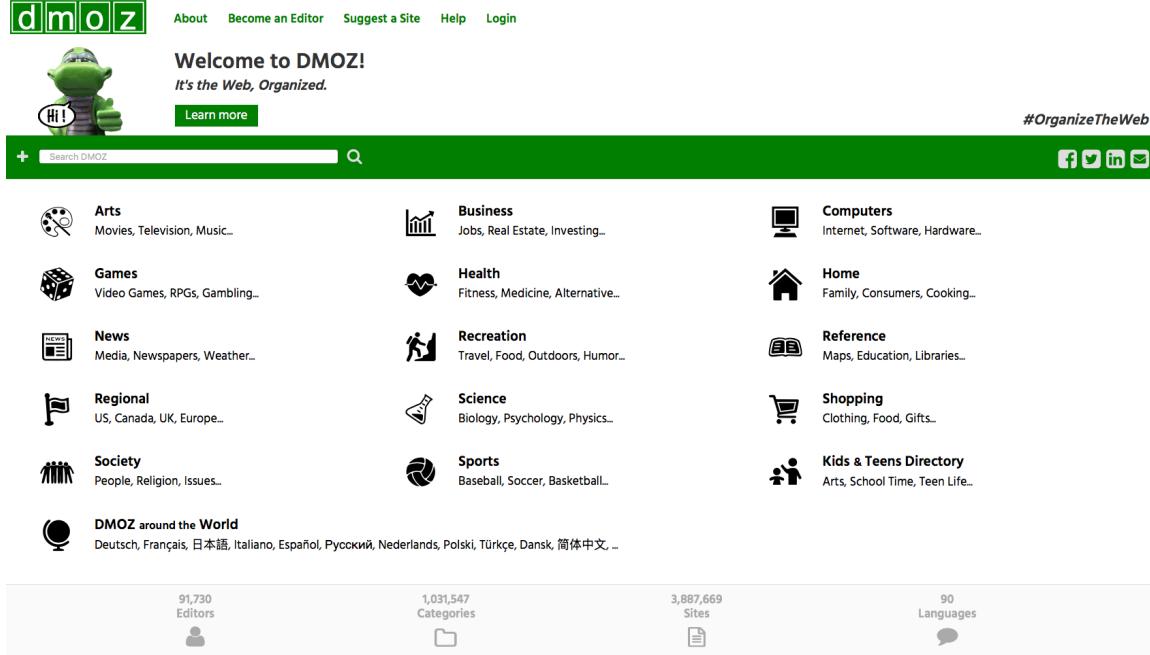


Broad Question

◆ How to organize the Web?

◆ First try: Human curated Web directories

➤ Yahoo, DMOZ, LookSmart



The screenshot shows the homepage of DMOZ (Open Directory Project). At the top, there's a navigation bar with links for About, Become an Editor, Suggest a Site, Help, and Login. Below this is a search bar with a magnifying glass icon and a "Learn more" button. The main content area features a grid of category icons and names. Each category has a small icon to its left and a brief description below it. The categories are:

- Arts: Movies, Television, Music...
- Business: Jobs, Real Estate, Investing...
- Computers: Internet, Software, Hardware...
- Games: Video Games, RPGs, Gambling...
- Health: Fitness, Medicine, Alternative...
- Home: Family, Consumers, Cooking...
- News: Media, Newspapers, Weather...
- Recreation: Travel, Food, Outdoors, Humor...
- Reference: Maps, Education, Libraries...
- Regional: US, Canada, UK, Europe...
- Science: Biology, Psychology, Physics...
- Shopping: Clothing, Food, Gifts...
- Society: People, Religion, Issues...
- Sports: Baseball, Soccer, Basketball...
- Kids & Teens Directory: Arts, School Time, Teen Life...
- DMOZ around the World: Deutsch, Français, 日本語, Italiano, Español, Русский, Nederlands, Polski, Türkçe, Dansk, 简体中文, ...

At the bottom of the page, there's a footer with statistics: 91,730 Editors, 1,031,547 Categories, 3,887,669 Sites, and 90 Languages. There's also a "Terms of Use" link.



The screenshot shows the Yahoo homepage from September 2016. The top navigation bar includes links for Headlines, Yahoo Lines, Info, and Add URL. A banner at the top says "How Open: Yahoo! Best Shop! Remove the Shopping Gap to Play". Below the banner is a search bar with a "Search" button and an "Options" link. The main content area is a list of categories under the heading "#OrganizeTheWeb". The categories are:

- Arts: Humor, Photography, Archives...
- Business and Economy: Directory, Investments, Classifieds, Taxes...
- Computers and Internet: Search, WWW, Software, Mathematics...
- Education: Universities, K-12, Colleges...
- Entertainment: TV, Movie, Music, Magazines...
- Government: Politics (Browz), Agencies, Law, Military...
- Health: Medicine, Drugs, Doctors, Fitness...
- News: York (Browz), Daily, Current Events...
- Recreation: Sports (Browz), Games, Travel, Avatars...
- Reference: Libraries, Dictionaries, Fact Books...
- Regional: Countries, Regions, U.S. States...
- Science: Cell Biology, Astronomy, Engineering...
- Social Science: Anthropology, Sociology, Economics...
- Society and Culture: People, Environment, Religion...

At the bottom right, there's a link for "Text-Only Yahoo - Contributions".

Broad Question (Cont'd)

◆ How to organize the Web?

◆ Second try: Web Search

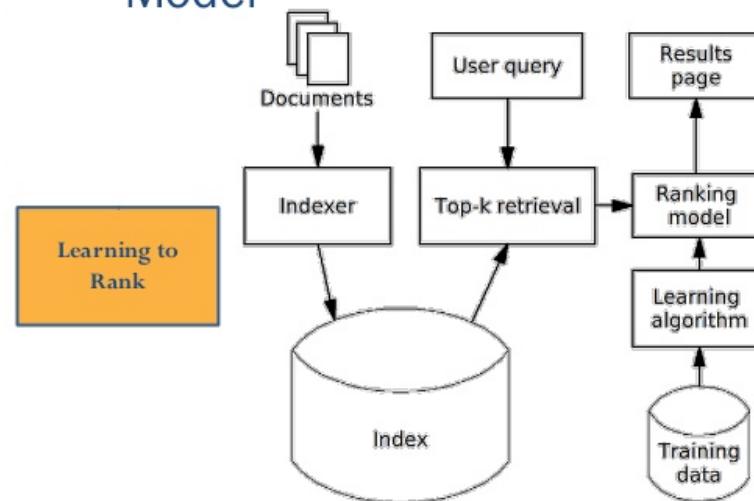
➤ **Information Retrieval** investigates:

Find relevant docs in a small
and trusted set

- Newspaper articles,
patents, etc.

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

Static Information Retrieval Model



Early Web Search

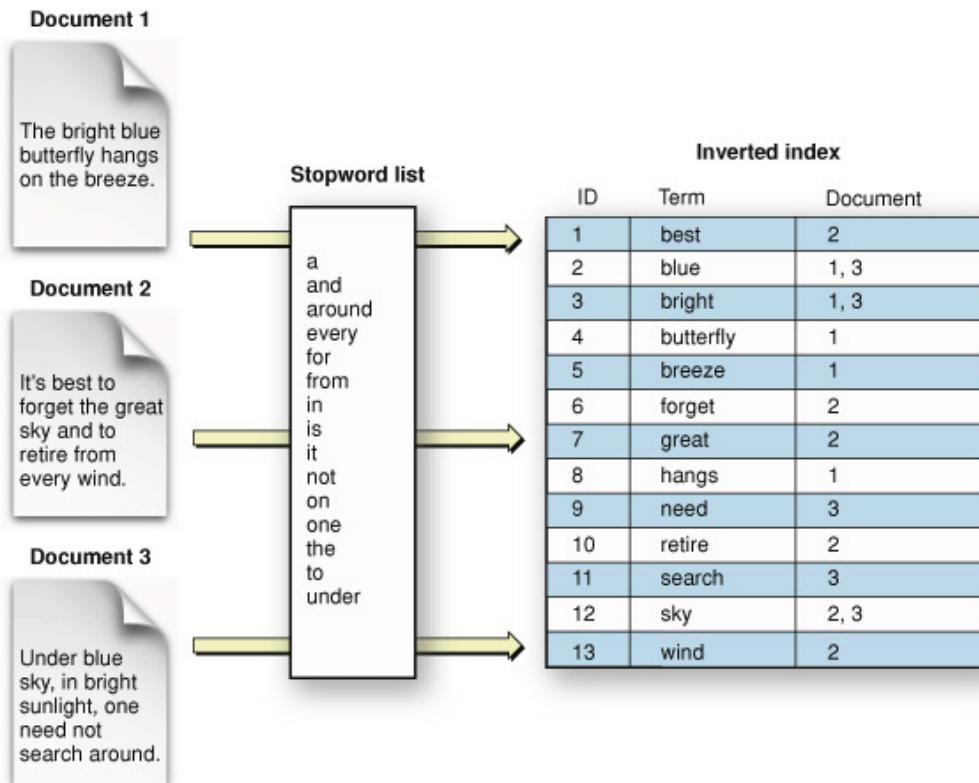
- ◆ Keywords extracted from web pages
 - E.g., title, content
 - Used to build **inverted index**

- ◆ Queries are matched with web pages
 - Via lookup in the **inverted index**
 - Pages ranked by occurrences of query keywords

```
"a": {2}  
"banana": {2}  
"is": {0, 1, 2}  
"it": {0, 1, 2}  
"what": {0, 1}
```

Inverted Index

- ◆ Problem: susceptible to **term spam**



Term Spam

- ◆ Disguise a page as something it is not about
 - E.g., adding thousands of keyword “movies”
 - Actual content may be some advertisement
 - Fool search engine to return it for query “movies”
- ◆ May even fade spam words into background
- ◆ Spam pages may be based on top-ranked pages

Web Search: Two Challenges

Two challenges of web search:

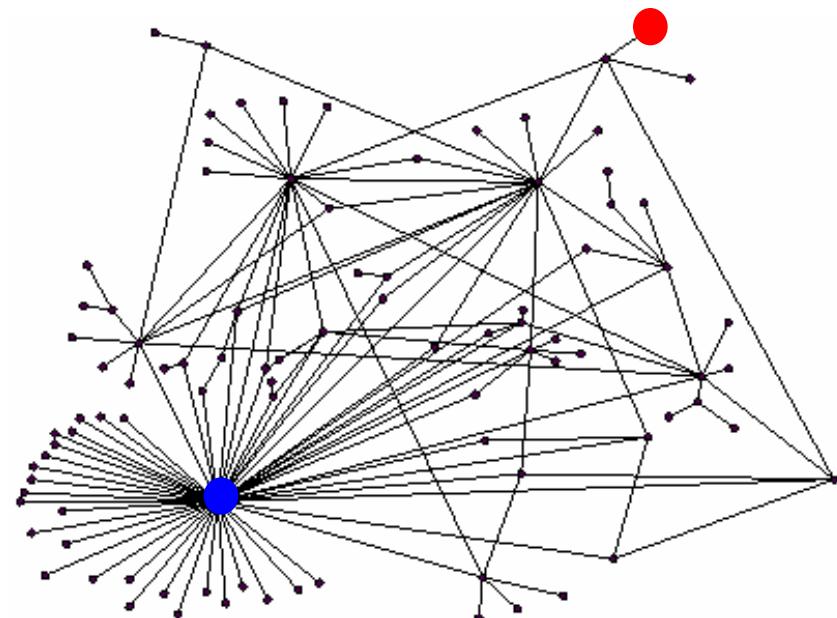
- ◆ (1) Web contains many sources of information
Who to “trust”?
 - Trick: Trustworthy pages may point to each other!
- ◆ (2) What is the “best” answer to query “newspaper”?
 - No single right answer
 - Trick: Pages that actually know about newspapers might all be pointing to many newspapers.

Ranking Nodes on the Graph

- ◆ All web pages are not equally “important”

www.joe-schmoe.com vs. www.usc.edu

- ◆ There is large diversity in the web-graph node connectivity
- ◆ Let's rank the pages by the link structure!



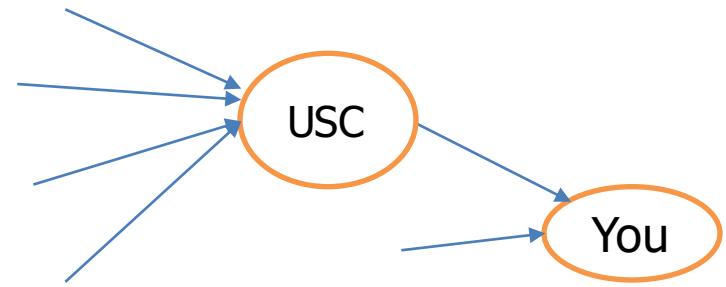
Link Analysis Algorithms

- ◆ We will cover the following **Link Analysis approaches** for computing **importance** of nodes in a graph:
 - Page Rank
 - Topic-Specific (Personalized) Page Rank
 - Web Spam Detection Algorithms.

PageRank: The “Flow” Formulation

PageRank: Combating Term Spam

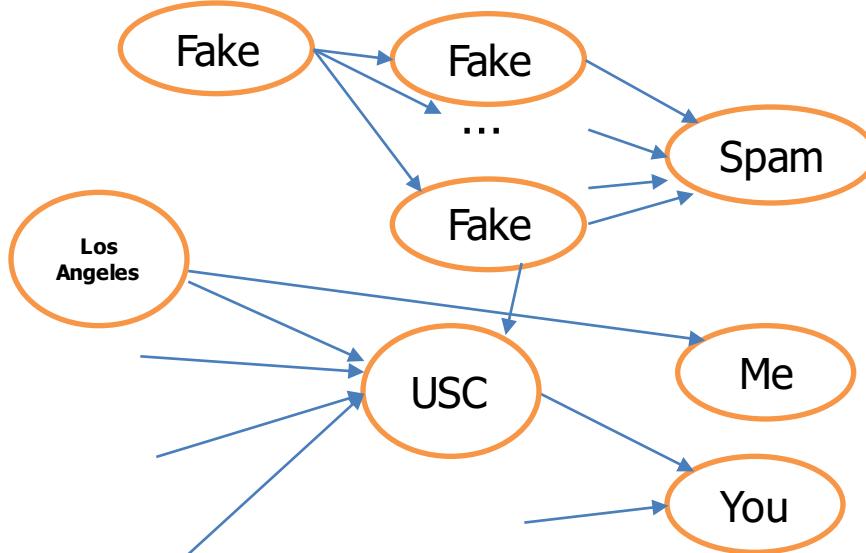
- ◆ Key idea: rank pages by linkage too
 - How **many** pages point to a page
 - How **important** these pages are=> PageRank



- ◆ USC.edu can be important
 - because many pages point to it
- ◆ Your home page can be important
 - If it is pointed to by USC ☺

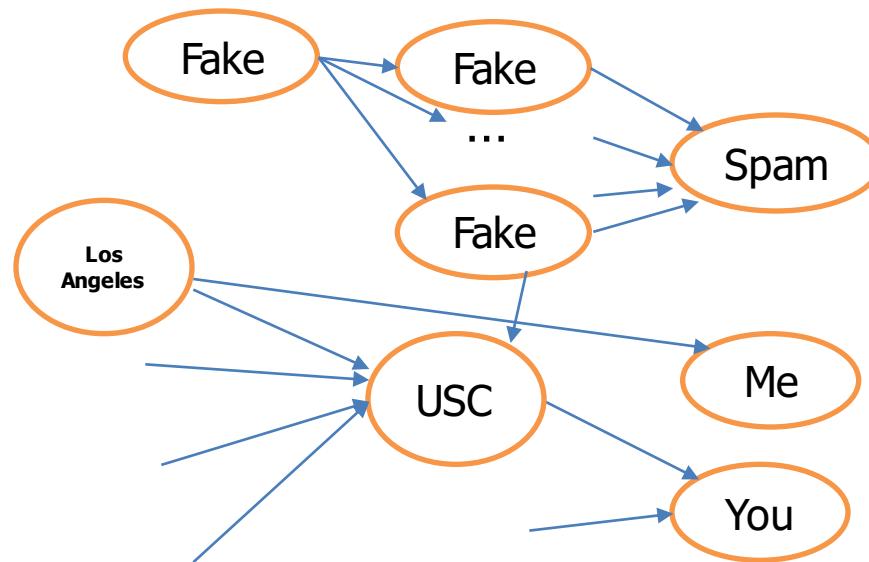
Random Surfer Model

- ◆ Random surfer of web
 - starts from any page
 - follows its outgoing links randomly
- ◆ Page is important if it attracts a large # of surfers



PageRank

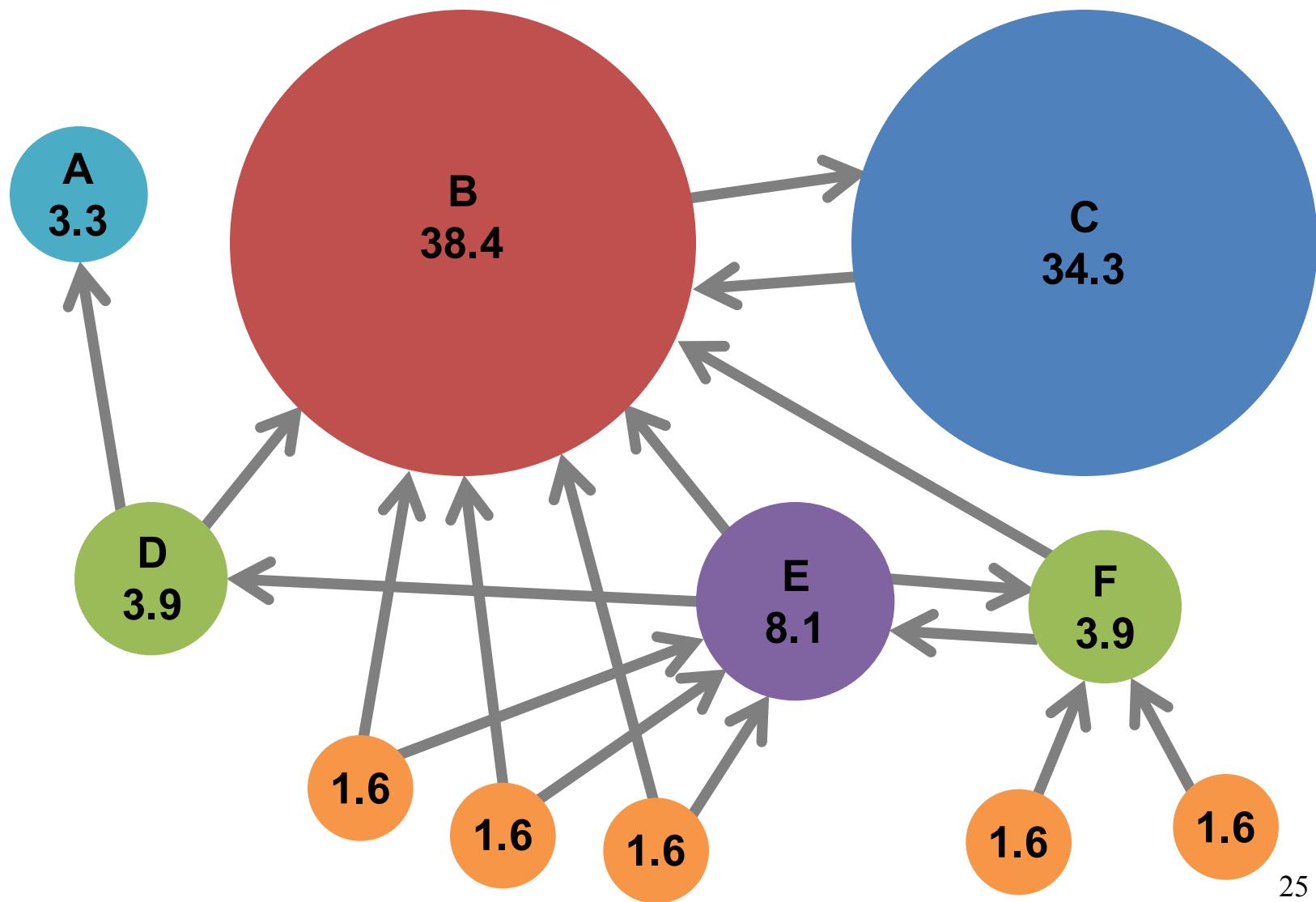
- ◆ Probability that a random surfer lands on the page



Intuition

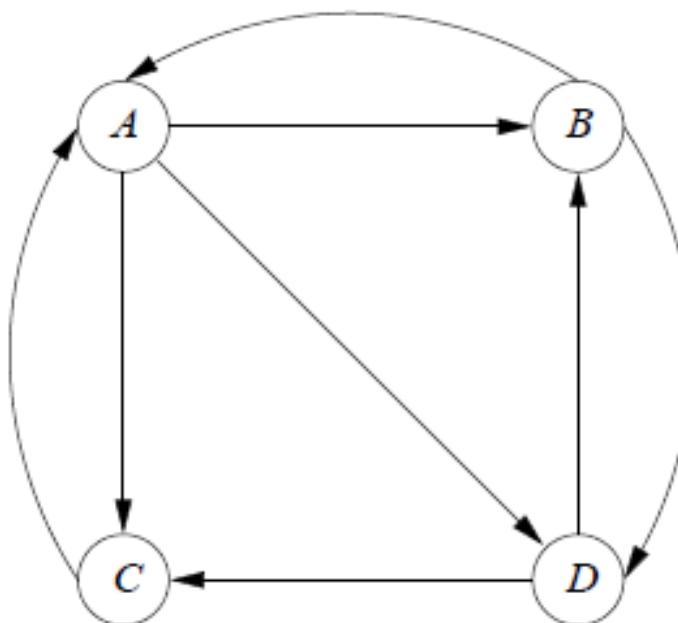
- ◆ If a page is important, then
 - many other pages may directly/indirectly link to it
 - random surfer can easily find it
- ◆ Spam pages are **less connected**
 - So less chance to attract random surfer
- ◆ Random surfer model more robust than manual approach
 - A **collective voting scheme.**

Example: PageRank Scores



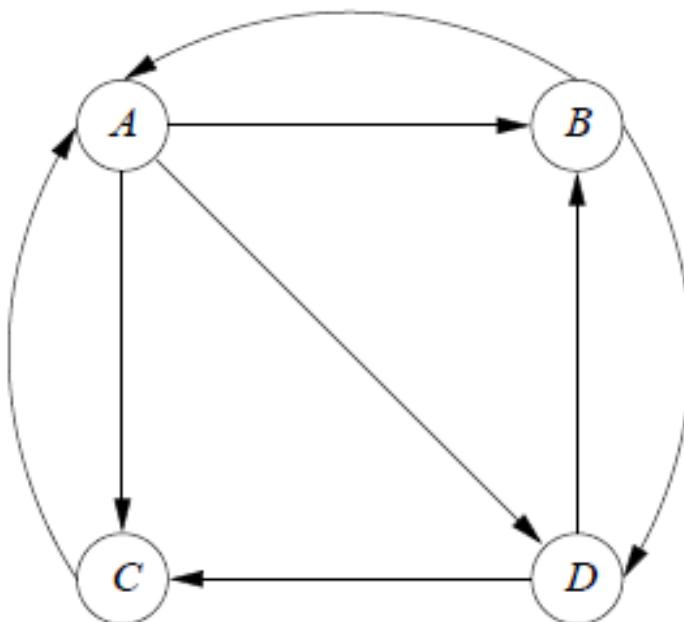
Assumption: A Strongly Connected Web Graph

- ◆ Nodes = pages
- ◆ Edges = hyperlinks between pages



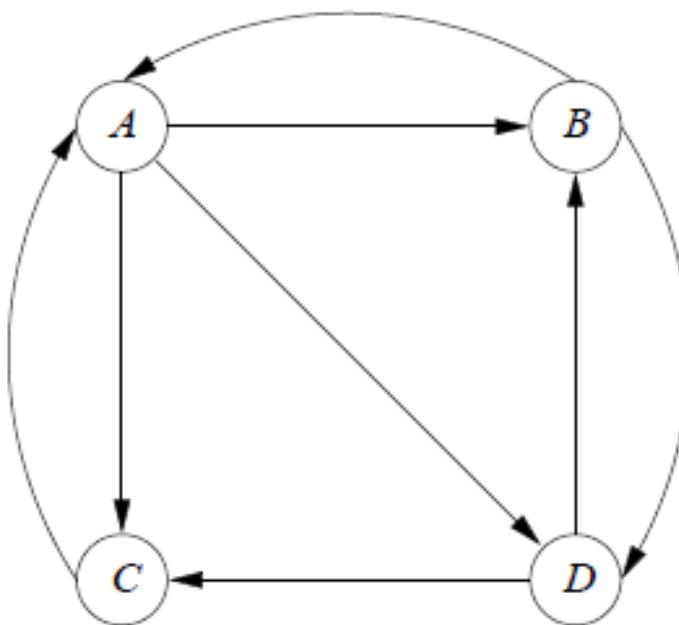
Model: Random Surfer on the Graph

- ◆ Can start at any node, say A
 - Can next go to B, C, or D, each with 1/3 prob.
 - If at B, can go to A and D, each with 1/2 prob.
 - So on...



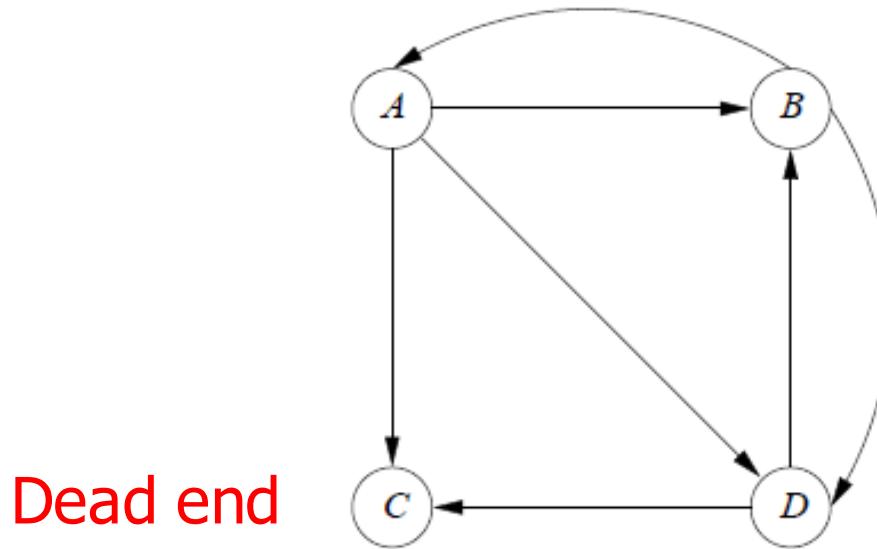
Random Surfer Property: Memoryless

- ◆ Where to go from node X is not affected by how the surfer got to X



Extreme Case: Dead End

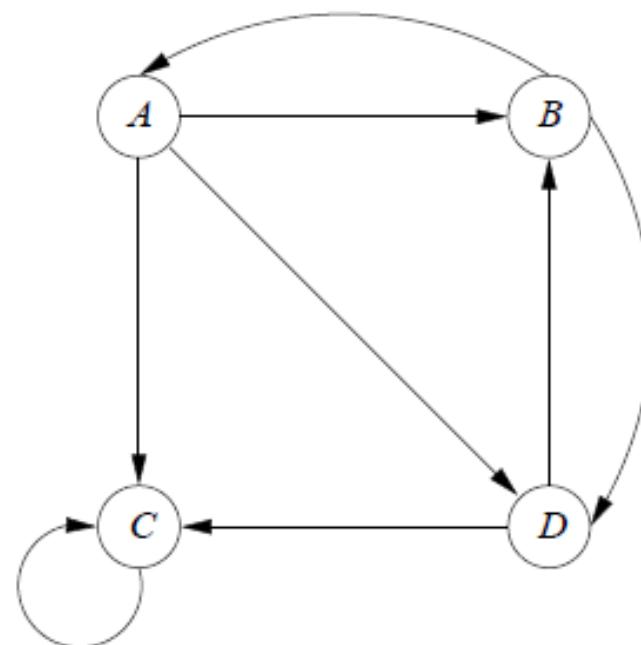
- ◆ Dead end: a page with no edges out
 - Absorb PageRanks
 - PageRank $\rightarrow 0$ for any page that can reach the dead end (including the dead end itself)



Extreme Case Spider Trap

- ◆ Group of pages with no edges going out of group
 - Absorb all PageRanks (rank of C → 1, others → 0)
 - Surfer can never leave, once trapped
 - Can have > 1 nodes

Spider trap



PageRank: Formulation Details

PageRank: Links as Votes

◆ 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:

- Eg:
- www.USC.edu ~ has 23,400 in-links
- www.joe-schmoe.com has 1 in-link

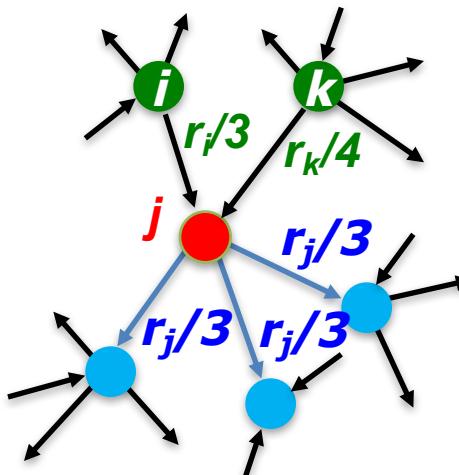
◆ Are all in-links are equal?

- Links from important pages count more
- Recursive question!

Simple Recursive Formulation

- ◆ 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

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

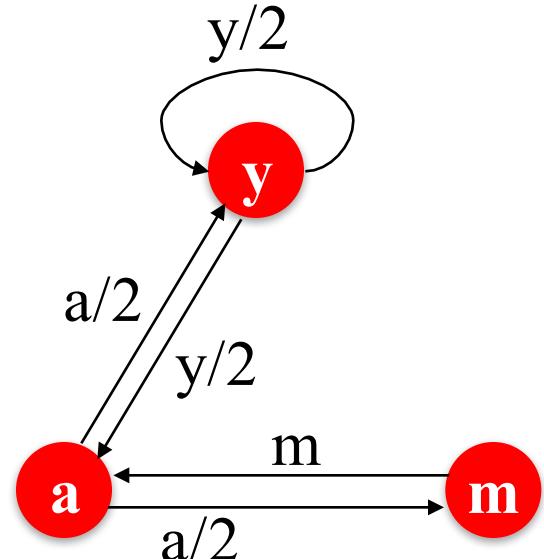


PageRank: The “Flow” Model

- ◆ A “vote” from an important page is worth more
- ◆ A page is important if it is pointed to by other important pages
- ◆ Define a “rank” r_j for page j

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

d_i = out-degree of node 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 equations

“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 scale factor

◆ Additional constraint forces uniqueness

$$r_y + r_m + r_a = 1$$

$$\text{➤ Solution: } r_y = \frac{2}{5}, \quad r_a = \frac{2}{5}, \quad r_m = \frac{1}{5}$$

◆ Gaussian elimination method works for small examples, but we need a better method for large web-size graphs

◆ We need a new formulation!

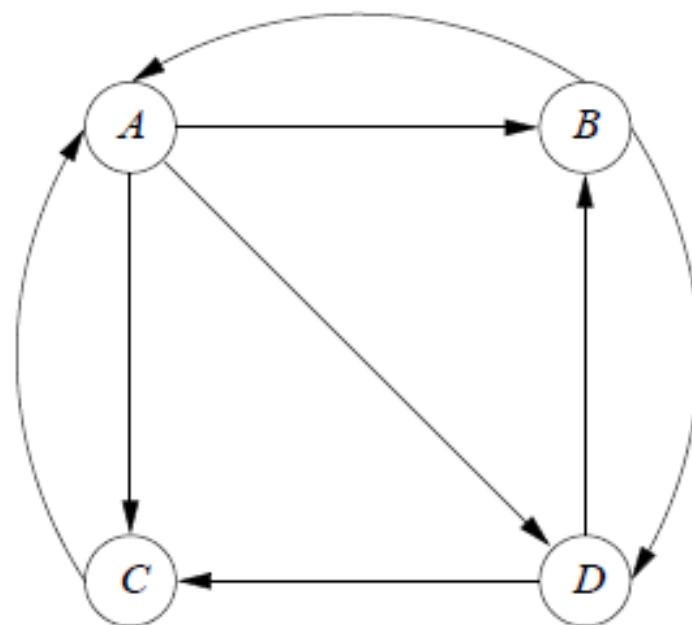
PageRank: Matrix formulation

- ◆ Stochastic Transition (or adjacency) Matrix M
- ◆ Suppose page j has n outlinks
 - If outlink $j \rightarrow i$, then $M_{ij} = 1/n$
 - Else $M_{ij} = 0$
- ◆ M is a column stochastic matrix
 - Columns sum to 1

Transition Matrix

- ◆ $M[i,j] = \text{prob. of going from node } j \text{ to node } i$
 - If j has k outgoing edges, prob. for each edge = $1/k$

$$M = \begin{matrix} & \begin{matrix} A & B & C & D \end{matrix} \\ \begin{matrix} A \\ B \\ C \\ D \end{matrix} & \begin{bmatrix} 0 & 1/2 & 1 & 0 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 1/2 & 0 & 0 \end{bmatrix} \end{matrix}$$



PageRank: Matrix formulation (Cont'd)

- ◆ Stochastic Transition (or adjacency) Matrix M
- ◆ Suppose page j has n outlinks
 - If outlink $j \rightarrow i$, then $M_{ij} = 1/n$
 - Else $M_{ij} = 0$
- ◆ M is a column stochastic matrix
 - Columns sum to 1
- ◆ Rank vector r is a vector with one entry per web page
 - r_i is the importance score of page i
- ◆ The flow equations can be written as $r = Mr$.

Example

- **Flow equation in matrix form: $\mathbf{M}\mathbf{r} = \mathbf{r}$**
- Suppose page j links to 3 pages, including i

$$\begin{matrix} & j \\ & | \\ i & \xrightarrow{\text{red arrow}} & \square \\ & \swarrow & \downarrow \\ & \square & \xrightarrow{\text{red arrow}} \\ & \swarrow & \downarrow \\ & \square & \xrightarrow{\text{red arrow}} \end{matrix} \quad \cdot \quad \begin{matrix} r_j \\ = \\ r_i \end{matrix} \quad = \quad \begin{matrix} r \end{matrix}$$

The diagram illustrates a 3x3 matrix M representing link flow between three pages. The columns are labeled i , j , and another unlabeled column. The rows are labeled i , j , and another unlabeled row. The entry in the i,j position is highlighted with a red arrow. The entry in the i,i position is labeled $1/3$ in green. The vector r has a red arrow pointing down its middle column, indicating it is a column vector. The result of the multiplication $M \cdot r$ is the vector r , which has a red arrow pointing down its middle column, indicating it is a column vector.

Stationary Distribution

- ◆ Limiting prob. distribution of random surfer
 - PageRanks are based on **limiting distribution**
 - **the probability destruction will converge eventually**
- ◆ Requirement for its existence
 - Graph is **strongly connected**: a node can reach any other node in the graph

=> Cannot have **dead ends, spider traps.**

Eigenvectors and Eigenvalues

- ◆ An **eigenvector** of a square matrix **A** is a non-zero vector **v** that, when the matrix multiples **v**, yields the same as when some scalar multiplies **v**, the scalar multiplier often being denoted by λ
- ◆ That is:

$$\mathbf{Av} = \lambda\mathbf{v}$$

- ◆ The number λ is called the **eigenvalue** of **A** corresponding to **v**.

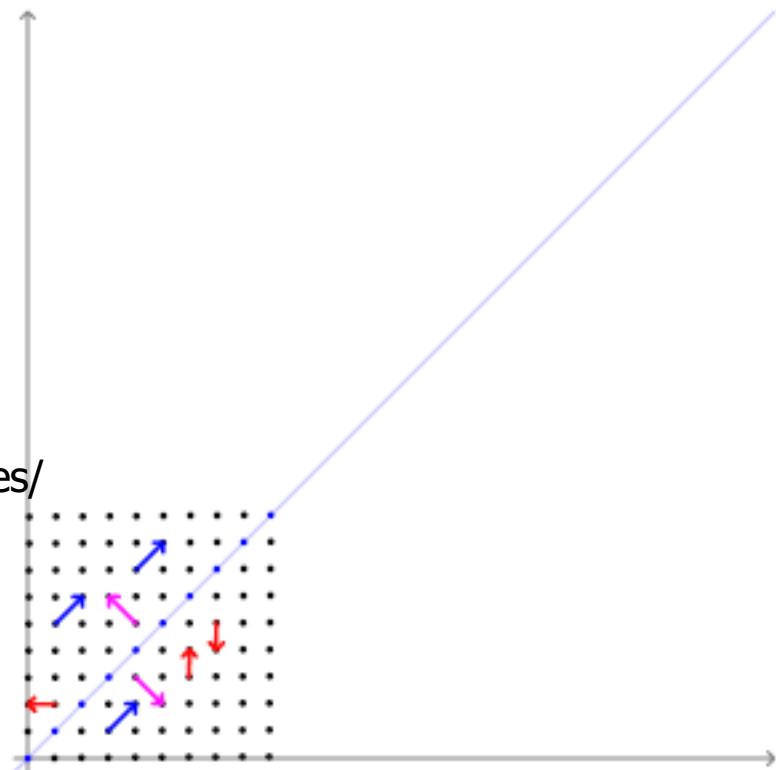
Eigenvalues and Eigenvectors Example

- ◆ The transformation matrix $A = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$ preserves the direction of vectors parallel to $\mathbf{v} = (1, -1)^T$ (in purple) and $\mathbf{w} = (1, 1)^T$ (in blue). The vectors in red are not parallel to either eigenvector, so, their directions are changed by the transformation.

$$Ax = \lambda x$$

<http://setosa.io/ev/eigenvectors-and-eigenvalues/>

https://en.wikipedia.org/wiki/Eigenvalues_and_eigenvectors



Eigenvector Formulation

- ◆ The flow equations can be written

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

limiting distribution

- ◆ So the rank vector \mathbf{r} is an eigenvector of the stochastic web matrix \mathbf{M}

- In fact, \mathbf{M} 's first or principal eigenvector, with corresponding eigenvalue 1
 - Largest eigenvalue of \mathbf{M} is 1 since \mathbf{M} is **column stochastic (with non-negative entries)**
 - We know \mathbf{r} is unit length and each column of \mathbf{M} sums to one

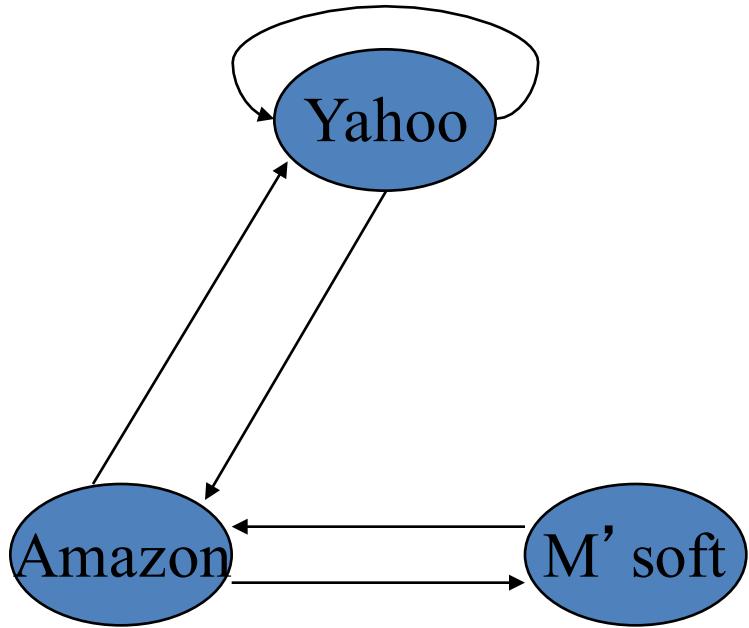
NOTE: \mathbf{x} is an eigenvector with the corresponding eigenvalue λ if:

$$\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$$

- ◆ We can now efficiently solve for \mathbf{r} !

- 1. Power Iteration:
https://en.wikipedia.org/wiki/Power_iteration
- 2. Use the principal eigenvector.

Example



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

$$r_a = r_y/2 + r_m$$

$$r_m = r_a/2$$

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

$$\mathbf{r} = \mathbf{Mr}$$

$$\begin{bmatrix} r_y \\ r_a \\ r_m \end{bmatrix} = \begin{bmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 0 & 1 \\ 0 & 1/2 & 0 \end{bmatrix} \begin{bmatrix} r_y \\ r_a \\ r_m \end{bmatrix}$$

Summary

- ◆ Web as a Directed Graph
 - Nodes: Webpages
 - Edges: Hyperlinks
- ◆ Early Web Search
 - Inverted Index
 - Term spam
- ◆ Link Analysis Algorithms
- ◆ Random Surfer Model
- ◆ PageRank: Links as Votes
- ◆ Eigenvalues and Eigenvectors.