CS249: ADVANCED DATA MINING

Graph and Network

Instructor: Yizhou Sun

yzsun@cs.ucla.edu

May 31, 2017

Methods Learnt

	Vector Data	Text Data	Recommender System	Graph & Network
Classification	**Decision Tree**; **Naïve Bayes**; **Logistic Regression** **SVM; NN**			Label Propagation
Clustering	**K-means; hierarchical clustering; DBSCAN; Mixture Models; kernel k-means**	**PLSA; LDA**	**Matrix Factorization**	SCAN; Spectral Clustering
Prediction	**Linear Regression** **GLM**		**Collaborative Filtering**	
Ranking				PageRank
Feature Representation		**Word embedding**		Network embedding

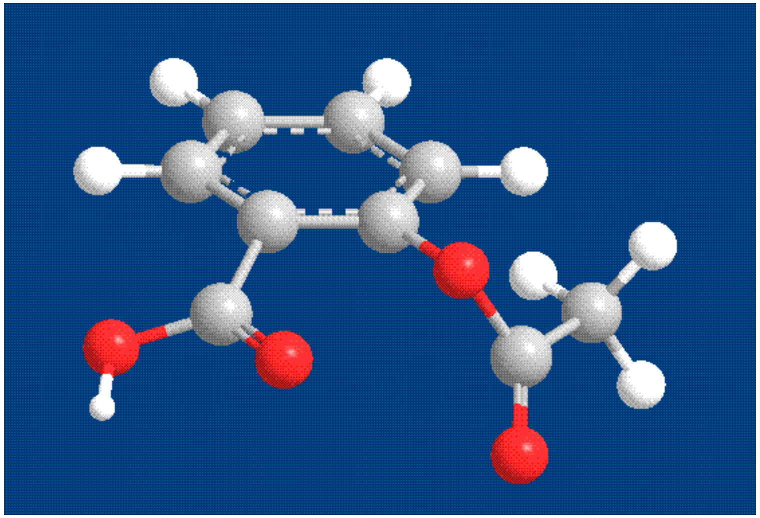
Methods to Learn

	Vector Data	Text Data	Recommender System	Graph & Network
Classification	Decision Tree; Naïve Bayes; Logistic Regression SVM; NN			Label Propagation
Clustering	K-means; hierarchical clustering; DBSCAN; Mixture Models; kernel k-means	PLSA; LDA	Matrix Factorization	SCAN; Spectral Clustering
Prediction	Linear Regression GLM		Collaborative Filtering	
Ranking				PageRank
Feature Representation		Word embedding		Network embedding

Mining Graph/Network Data

- Introduction to Graph/Network Data
- PageRank
- Classification via Label Propagation
- Spectral Clustering
- Summary

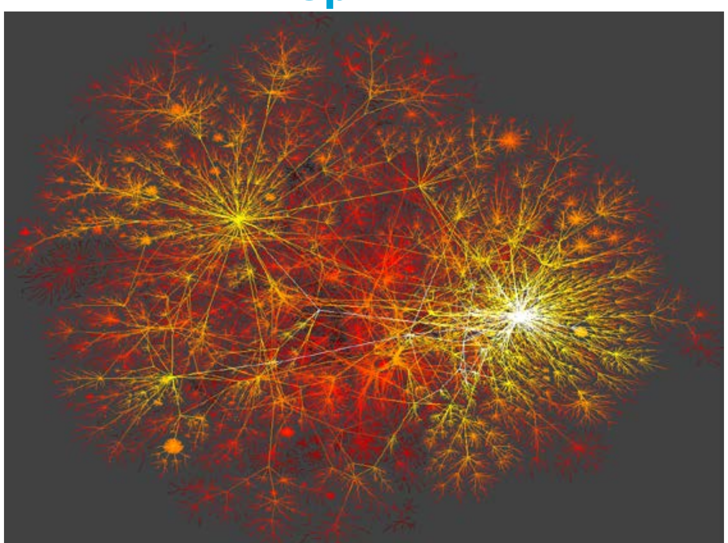
Graph, Graph, Everywhere



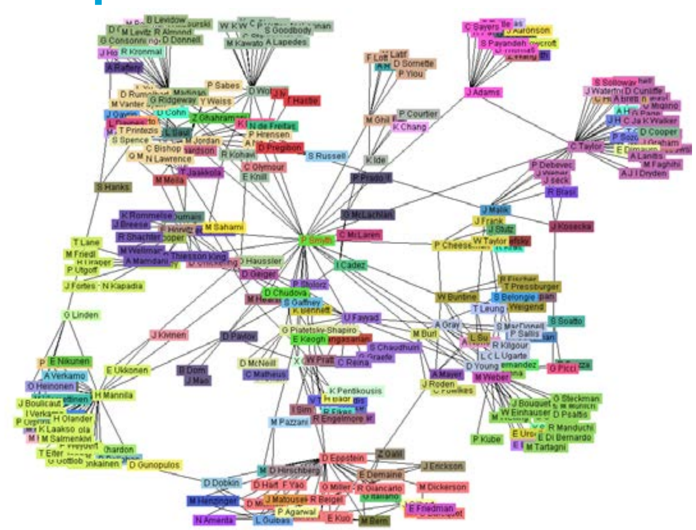
Aspirin

from H. Jeong et al Nature 411, 41 (2001)

Yeast protein interaction network



Internet



Co-author network

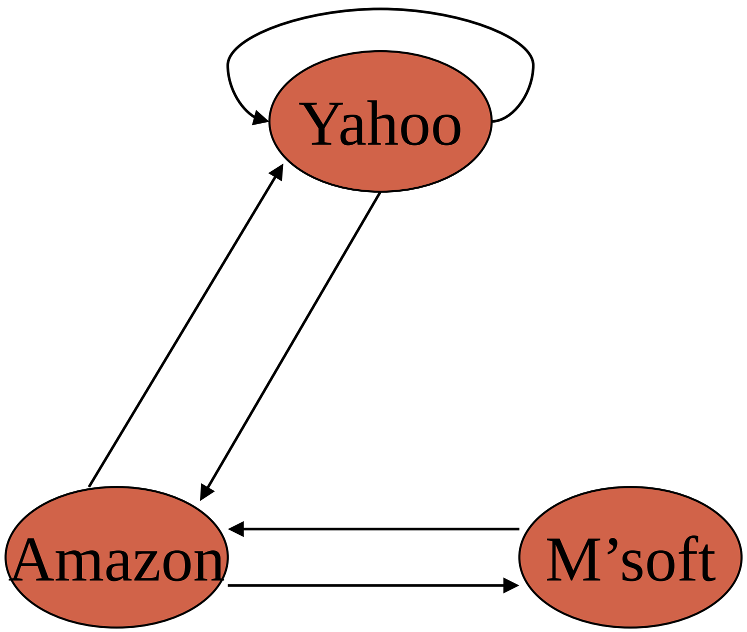
Why Graph Mining?

- Graphs are ubiquitous
  - Chemical compounds (Cheminformatics)
  - Protein structures, biological pathways/networks (Bioinformactics)
  - Program control flow, traffic flow, and workflow analysis
  - XML databases, Web, and social network analysis
- Graph is a general model
  - Trees, lattices, sequences, and items are degenerated graphs
- Diversity of graphs
  - Directed vs. undirected, labeled vs. unlabeled (edges & vertices), weighted, with angles & geometry (topological vs. 2-D/3-D)
- Complexity of algorithms: many problems are of high complexity

Representation of a Graph

- $G =< V, E >$
  - $V = \{u_1, \dots, u_n\}$: node set
  - $E \subseteq V \times V$: edge set
- Adjacency matrix
  - $A = \{a_{ij}\}, i, j = 1, \dots, N$
    - $a_{ij} = 1, if \ < u_i, u_j > \in E$
    - $a_{ij} = 0, if \ < u_i, u_j > \notin E$
  - Undirected graph vs. Directed graph
    - $A = A^{\mathrm{T}}\ vs. A \neq A^{\mathrm{T}}$
  - Weighted graph
    - Use *W* instead of *A*, where w_{ij} represents the weight of edge $< u_i, u_j >$

Example



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

Adjacency matrix A

Mining Graph/Network Data

- Introduction to Graph/Network Data
- PageRank
- Classification via Label Propagation
- Spectral Clustering
- Summary

The History of PageRank

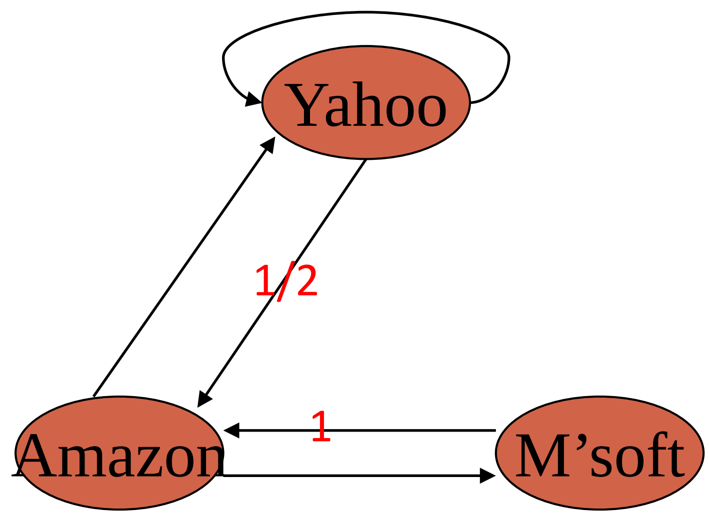
- PageRank was developed by Larry Page (hence the name *Page*-Rank) and Sergey Brin.
- It is first as part of a research project about a new kind of search engine. That project started in 1995 and led to a functional prototype in 1998.
- Shortly after, Page and Brin founded Google.

Ranking web pages

- Web pages are not equally "important"
  - www.cnn.com vs. a personal webpage
- Inlinks as votes
  - The more inlinks, the more important
- Are all inlinks equal?
  - Higher ranked inlink should play a more important role
  - Recursive question!

Simple recursive formulation

- Each link's vote is proportional to the importance of its source page
- If page P with importance x has n outlinks, each link gets x/n votes
- Page P's own importance is the sum of the votes on its inlinks



Matrix formulation

- Matrix **M** has one row and one column for each web page
- Suppose page j has n outlinks
  - If j -> i, then $M_{ij}=1/n$
  - Else $M_{ij}=0$

$$\begin{array}{cccc} & y & a & m \\ y & 1 & 1 & 0 \\ a & 1 & 0 & 1 \\ m & 0 & 1 & 0 \end{array} \Rightarrow \boxed{½, 0, 1}$$

- **M** is a column stochastic matrix
  - Columns sum to 1
- Suppose **r** is a vector with one entry per web page
  - r_i is the importance score of page i
  - Call it the rank vector
  - $|\mathbf{r}| = 1$ (i.e., $r_1 + r_2 + \cdots + r_N = 1$)

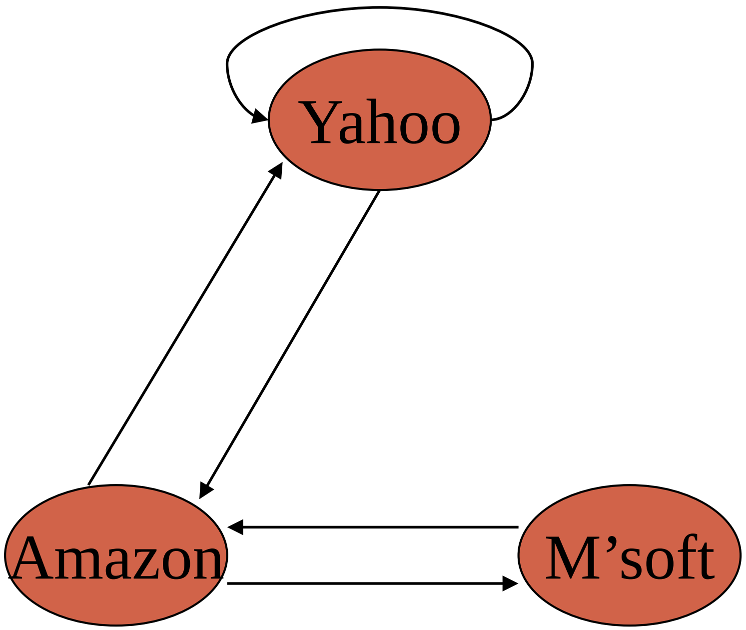
Eigenvector formulation

- The flow equations can be written

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

- So the rank vector is an eigenvector of the stochastic web matrix
  - In fact, its first or principal eigenvector, with corresponding eigenvalue 1

Example



$$y = y/2 + a/2$$

$$a = y/2 + m$$

$$m = a/2$$

$$\begin{array}{cccc} & y & a & m \\ \begin{matrix} y \\ a \\ m \end{matrix} & \multicolumn{3}{c}{\boxed{\begin{matrix} 1/2 & 1/2 & 0 \\ 1/2 & 0 & 1 \\ 0 & 1/2 & 0 \end{matrix}}} \end{array}$$

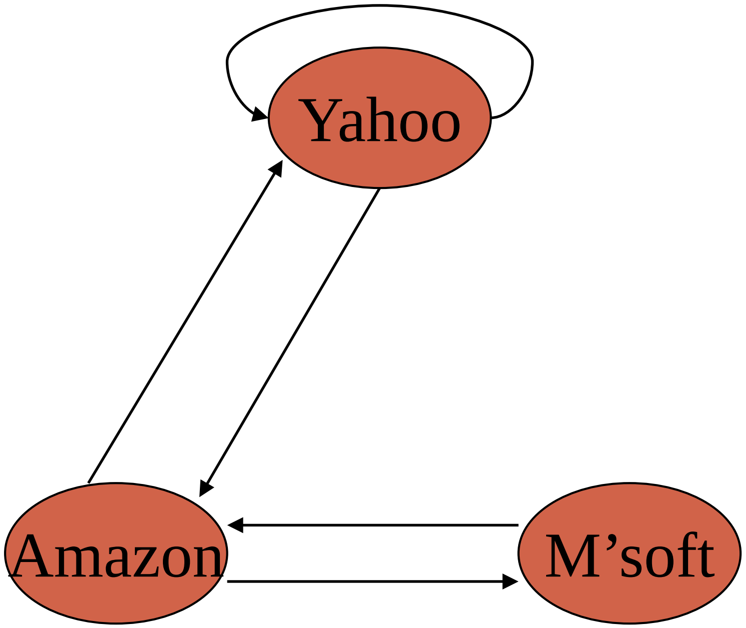
r = M * r

$$\boxed{\begin{matrix} y \\ a \\ m \end{matrix}} = \boxed{\begin{matrix} 1/2 & 1/2 & 0 \\ 1/2 & 0 & 1 \\ 0 & 1/2 & 0 \end{matrix}} \boxed{\begin{matrix} y \\ a \\ m \end{matrix}}$$

Power Iteration method

- Simple iterative scheme
- Suppose there are N web pages
  - Initialize: $\mathbf{r}^0 = [1/N,\ldots.,1/N]^T$
  - Iterate: $\mathbf{r}^{k+1} = \mathbf{M}\mathbf{r}^k$
  - Stop when $|\mathbf{r}^{k+1} - \mathbf{r}^k|_1 < \varepsilon$
    - $|\mathbf{x}|_1 = \sum_{1\le i\le N}|x_i|$ is the L_1 norm
    - Can use any other vector norm e.g., Euclidean

Power Iteration Example



$$\begin{array}{c|ccc|} & y & a & m \\ \hline y & 1/2 & 1/2 & 0 \\ a & 1/2 & 0 & 1 \\ m & 0 & 1/2 & 0 \\ \hline \end{array}$$

$$\begin{array}{c} y \\ a \\ m \end{array} = \begin{array}{cccccc} 1/3 & 1/3 & 5/12 & 3/8 & & 2/5 \\ 1/3 & 1/2 & 1/3 & 11/24 & \cdots & 2/5 \\ 1/3 & 1/6 & 1/4 & 1/6 & & 1/5 \\ \\ \boldsymbol{r}_0 & \boldsymbol{r}_1 & \boldsymbol{r}_2 & \boldsymbol{r}_3 & \ldots & \boldsymbol{r}^* \end{array}$$

Random Walk Interpretation

- Imagine a random web surfer
  - At any time t, surfer is on some page P
  - At time t+1, the surfer follows an outlink from P uniformly at random
  - Ends up on some page Q linked from P
  - Process repeats indefinitely
- Let **p**(t) be a vector whose i^{th} component is the probability that the surfer is at page i at time t
  - **p**(t) is a probability distribution on pages

The stationary distribution

- Where is the surfer at time t+1?
  - Follows a link uniformly at random
  - **p**(t+1) = **Mp**(t)
- Suppose the random walk reaches a state such that **p**(t+1) = **Mp**(t) = **p**(t)
  - Then **p**(t) is called a stationary distribution for the random walk
- Our rank vector **r** satisfies **r** = **Mr**
  - So it is a stationary distribution for the random surfer

Existence and Uniqueness

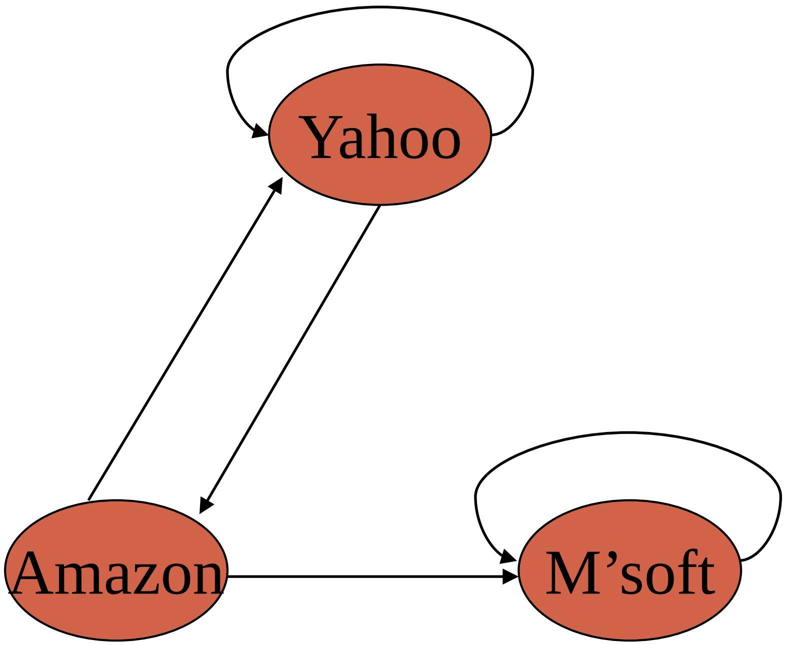
A central result from the theory of random walks (aka Markov processes):

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.

Spider traps

- A group of pages is a spider trap if there are no links from within the group to outside the group
  - Random surfer gets trapped
- Spider traps violate the conditions needed for the random walk theorem

Microsoft becomes a spider trap



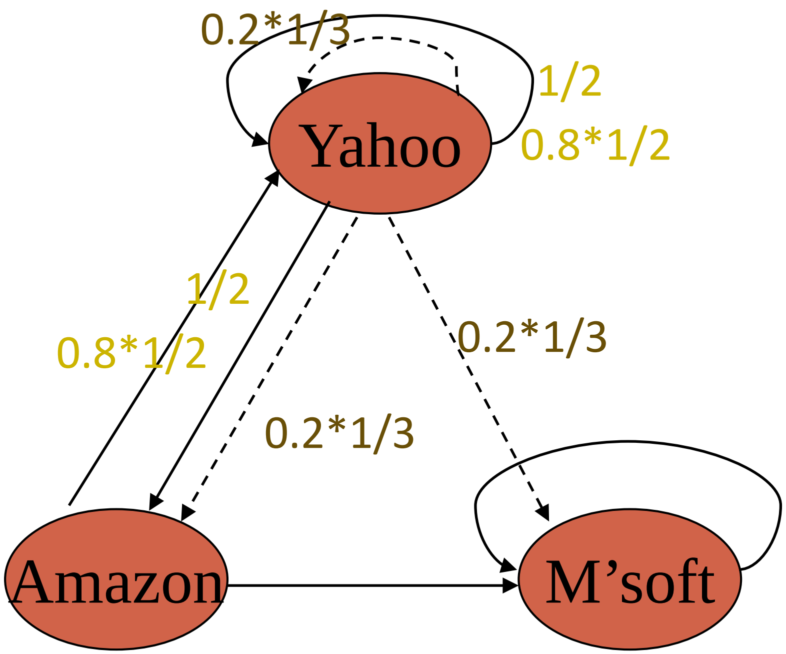
$$\begin{array}{c|ccc|} & y & a & m \\ \hline y & 1/2 & 1/2 & 0 \\ a & 1/2 & 0 & 0 \\ m & 0 & 1/2 & 1 \\ \hline \end{array}$$

$$\begin{array}{lllllll} y & & 1/3 & 1/3 & 1/4 & 5/24 & & 0 \\ a & = & 1/3 & 1/6 & 1/6 & 1/8 & \ldots & 0 \\ m & & 1/3 & 1/2 & 7/12 & 2/3 & & 1 \end{array}$$

Random teleports

- 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 of spider trap within a few time steps

Random teleports ($\beta = 0.8$)



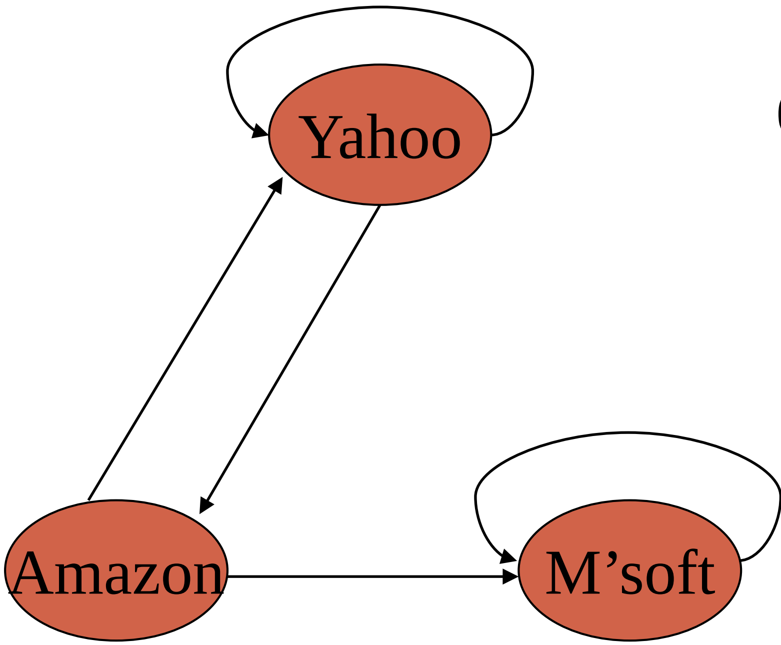
: teleport links from “Yahoo”

$$
\begin{array}{c|c|}
 & y \\
\hline
y & 1/2 \\
a & 1/2 \\
m & 0 \\
\hline
\end{array}
\;=\;
0.8 * \begin{array}{|c|}
y \\
\hline
1/2 \\
1/2 \\
0 \\
\hline
\end{array}
\;+\; 0.2 * \begin{array}{|c|}
y \\
\hline
1/3 \\
1/3 \\
1/3 \\
\hline
\end{array}
$$

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

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

Random teleports ($\beta = 0.8$)



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

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

$$\begin{array}{c} y \\ a \\ m \end{array} = \begin{bmatrix} 0.333 \\ 0.333 \\ 0.333 \end{bmatrix} \begin{bmatrix} 0.333 \\ 0.200 \\ 0.467 \end{bmatrix} \begin{bmatrix} 0.280 \\ 0.200 \\ 0.520 \end{bmatrix} \begin{bmatrix} 0.259 \\ 0.179 \\ 0.563 \end{bmatrix} \ldots \begin{bmatrix} 7/33 \\ 5/33 \\ 21/33 \end{bmatrix}$$

Matrix formulation

- Suppose there are N pages
  - Consider a page j, with set of outlinks O(j)
  - We have $M_{ij} = 1/|O(j)|$ when j->i and $M_{ij} = 0$ otherwise
  - The random teleport is equivalent to
    - adding a teleport link from j to every other page with probability $(1-\beta)/N$
    - reducing the probability of following each outlink from $1/|O(j)|$ to $\beta/|O(j)|$
    - Equivalent: tax each page a fraction $(1-\beta)$ of its score and redistribute evenly

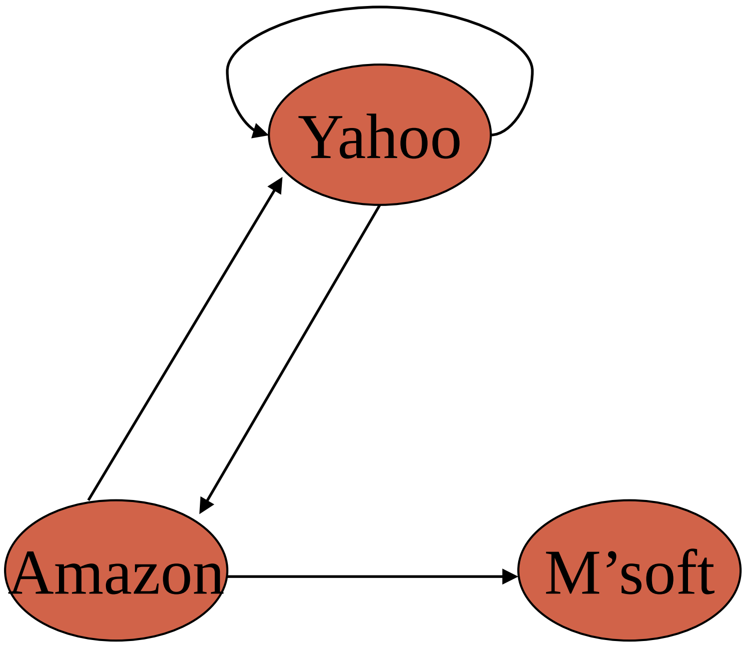
PageRank

- Construct the N-by-N matrix **A** as follows
  - $A_{ij} = \beta M_{ij} + (1-\beta)/N$
- Verify that **A** is a stochastic matrix
- The page rank vector **r** is the principal eigenvector of this matrix
  - satisfying $\mathbf{r} = \mathbf{A}\mathbf{r}$
- Equivalently, **r** is the stationary distribution of the random walk with teleports

Dead ends

- Pages with no outlinks are "dead ends" for the random surfer
  - Nowhere to go on next step

Microsoft becomes a dead end



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

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

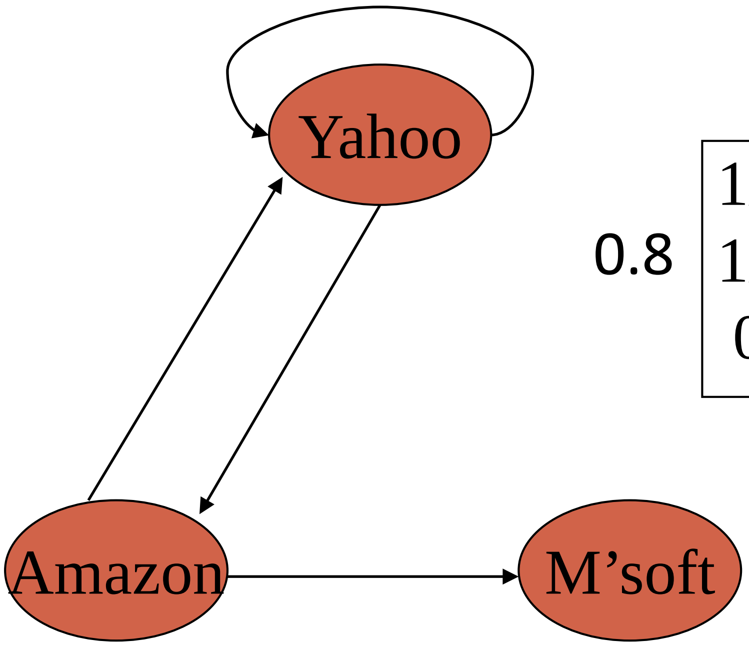
Non-stochastic!

$$\begin{array}{c} y \\ a \\ m \end{array} = \begin{array}{c} 1/3 \\ 1/3 \\ 1/3 \end{array} \quad \begin{array}{c} 1/3 \\ 0.2 \\ 0.2 \end{array} \quad \cdots \quad \begin{array}{c} 0 \\ 0 \\ 0 \end{array}$$

Dealing with dead-ends

- Teleport
  - Follow random teleport links with probability 1.0 from dead-ends
  - Adjust matrix accordingly
- Prune and propagate
  - Preprocess the graph to eliminate dead-ends
    - Might require multiple passes
  - Compute page rank on reduced graph
  - Approximate values for deadends by propagating values from reduced graph

Dealing dead end: teleport

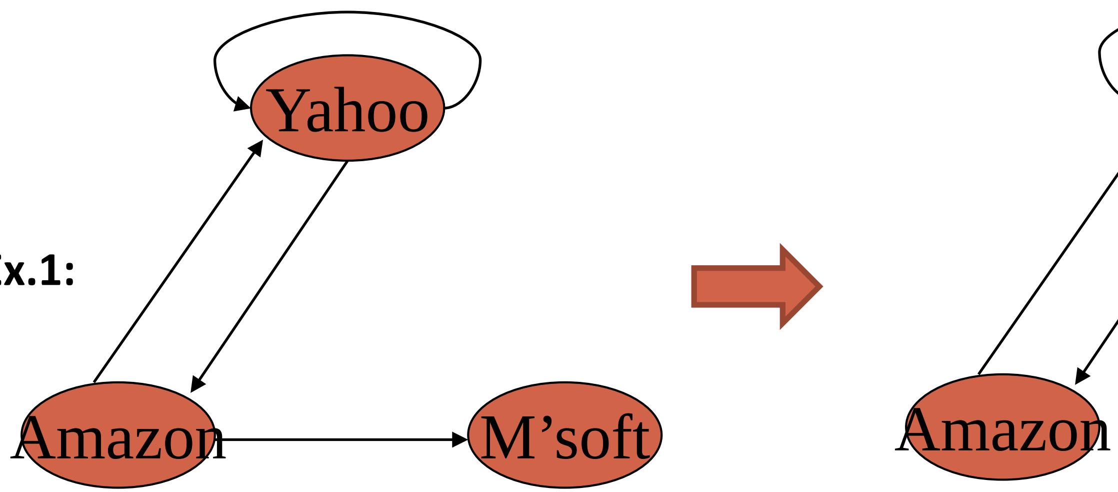


$$0.8 \begin{bmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 0 & 0 \\ 0 & 1/2 & 0 \end{bmatrix} + \begin{bmatrix} 0.2*1/3 & 0.2*1/3 & 1*1/3 \\ 0.2*1/3 & 0.2*1/3 & 1*1/3 \\ 0.2*1/3 & 0.2*1/3 & 1*1/3 \end{bmatrix}$$

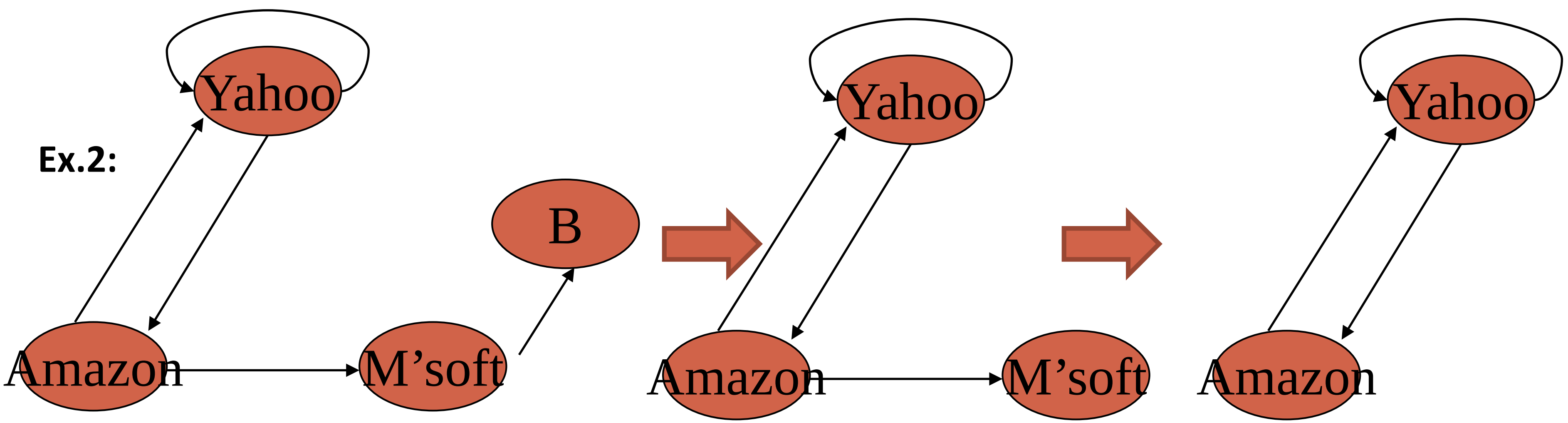
$$\begin{array}{c} y \\ a \\ m \end{array} \begin{bmatrix} 7/15 & 7/15 & 1/3 \\ 7/15 & 1/15 & 1/3 \\ 1/15 & 7/15 & 1/3 \end{bmatrix}$$

Dealing dead end: reduce graph

Ex.1:



Ex.2:



Computing PageRank

- Key step is matrix-vector multiplication
  - $\mathbf{r}^{\text{new}} = \mathbf{A}\mathbf{r}^{\text{old}}$
- Easy if we have enough main memory to hold **A**, $\mathbf{r}^{\text{old}}$, $\mathbf{r}^{\text{new}}$
- Say N = 1 billion pages
  - We need 4 bytes for each entry (say)
  - 2 billion entries for vectors, approx 8GB
  - Matrix A has N^2 entries
    - 10^{18} is a large number!

Rearranging the equation

$\mathbf{r} = \mathbf{Ar}$, where

$A_{ij} = \beta M_{ij} + (1-\beta)/N$

$r_i = \sum_{1 \le j \le N} A_{ij}\, r_j$

$r_i = \sum_{1 \le j \le N} [\beta M_{ij} + (1-\beta)/N]\, r_j$

$\quad = \beta \sum_{1 \le j \le N} M_{ij}\, r_j + (1-\beta)/N \sum_{1 \le j \le N} r_j$

$\quad = \beta \sum_{1 \le j \le N} M_{ij}\, r_j + (1-\beta)/N$, since $|\mathbf{r}| = 1$

$\mathbf{r} = \beta \mathbf{Mr} + [(1-\beta)/N]_N$

where $[x]_N$ is an N-vector with all entries x

Sparse matrix formulation

- We can rearrange the page rank equation:
  - $\mathbf{r} = \beta\mathbf{M}\mathbf{r} + [(1-\beta)/N]_N$
  - $[(1-\beta)/N]_N$ is an N-vector with all entries $(1-\beta)/N$
- **M** is a sparse matrix!
  - 10 links per node, approx 10N entries
- So in each iteration, we need to:
  - Compute $\mathbf{r}^{new} = \beta\mathbf{M}\mathbf{r}^{old}$
  - Add a constant value $(1-\beta)/N$ to each entry in $\mathbf{r}^{new}$

Sparse matrix encoding

- Encode sparse matrix using only nonzero entries
  - Space proportional roughly to number of links
  - say 10N, or 4*10*1 billion = 40GB
  - still won't fit in memory, but will fit on disk

source node	degree	destination nodes
0	3	1, 5, 7
1	5	17, 64, 113, 117, 245
2	2	13, 23

Basic Algorithm

- Assume we have enough RAM to fit $\mathbf{r}^{new}$, plus some working memory
  - Store $\mathbf{r}^{old}$ and matrix $\mathbf{M}$ on disk

Basic Algorithm:

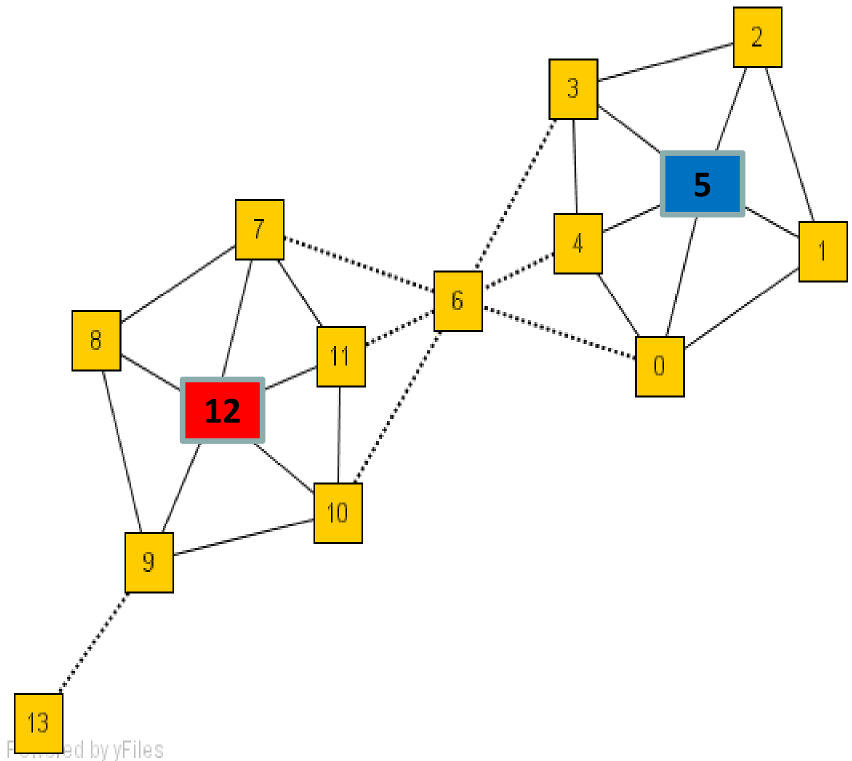
- Initialize: $\mathbf{r}^{old} = [1/N]_N$
- Iterate:
  - Update: Perform a sequential scan of $\mathbf{M}$ and $\mathbf{r}^{old}$ to update $\mathbf{r}^{new}$
  - Write out $\mathbf{r}^{new}$ to disk as $\mathbf{r}^{old}$ for next iteration
  - Every few iterations, compute $|\mathbf{r}^{new}-\mathbf{r}^{old}|$ and stop if it is below threshold
    - Need to read in both vectors into memory

Mining Graph/Network Data

- Introduction to Graph/Network Data
- PageRank
- Classification via Label Propagation
- Spectral Clustering
- Summary

Label Propagation in the Network

- Given a network, some nodes are given labels, can we classify the unlabeled nodes by using link information?
  - E.g., Node 12 belongs to Class 1 and Node 5 Belongs to Class 2



Reference

- Learning from Labeled and Unlabeled Data with Label Propagation
  - By Xiaojin Zhu and Zoubin Ghahramani
  - http://www.cs.cmu.edu/~zhuxj/pub/CMU-CALD-02-107.pdf

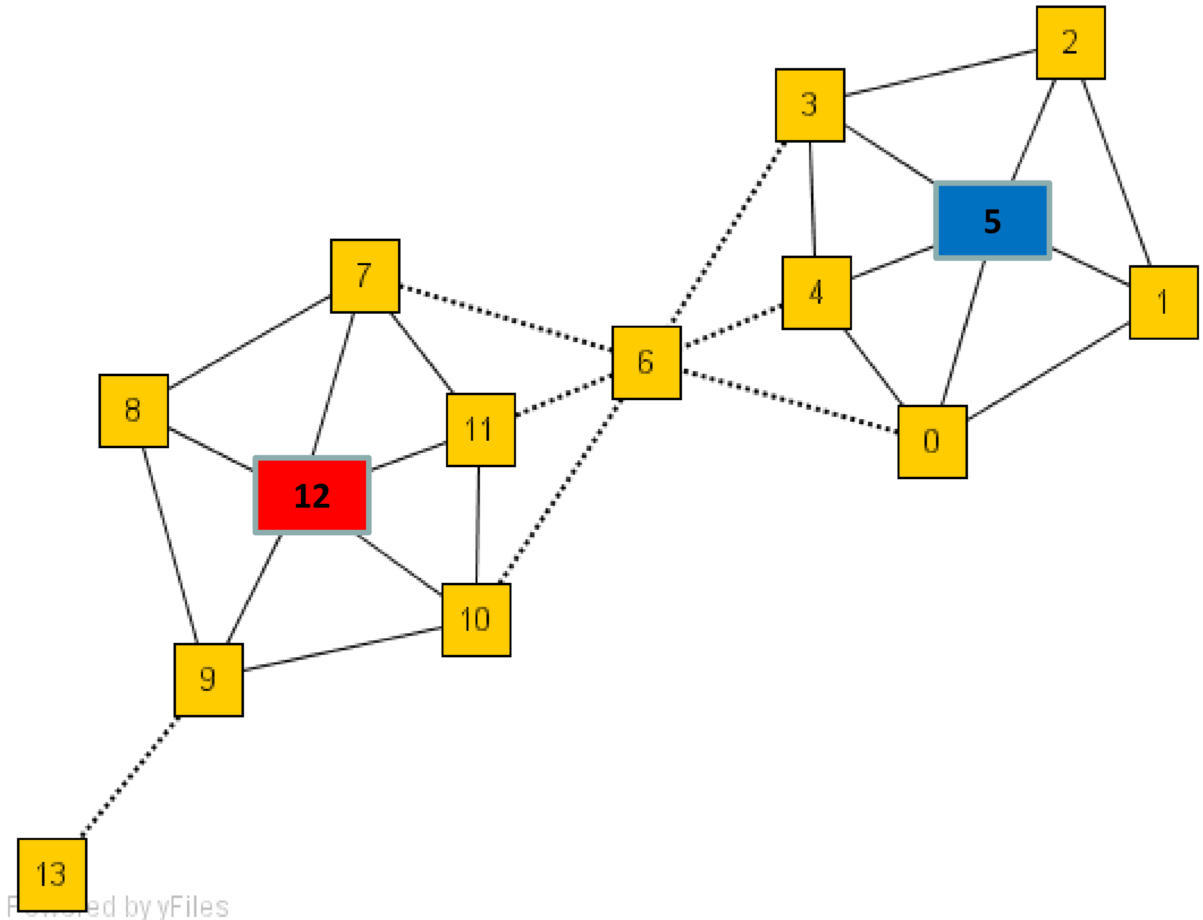
Problem Formalization

- Given n nodes
  - l with labels (e.g., $Y_1, Y_2, \dots, Y_l$ *are known*)
  - u without labels (e.g., $Y_{l+1}, Y_{l+2}, \dots, Y_n$ are unknown)
  - Y *is the* $n \times C$ *label matrix*
    - C is the number of labels (classes)
- The adjacency matrix is W
- The probabilistic transition matrix T
  - $T_{ij} = P(j \rightarrow i) = \frac{w_{ij}}{\sum_k w_{kj}}$

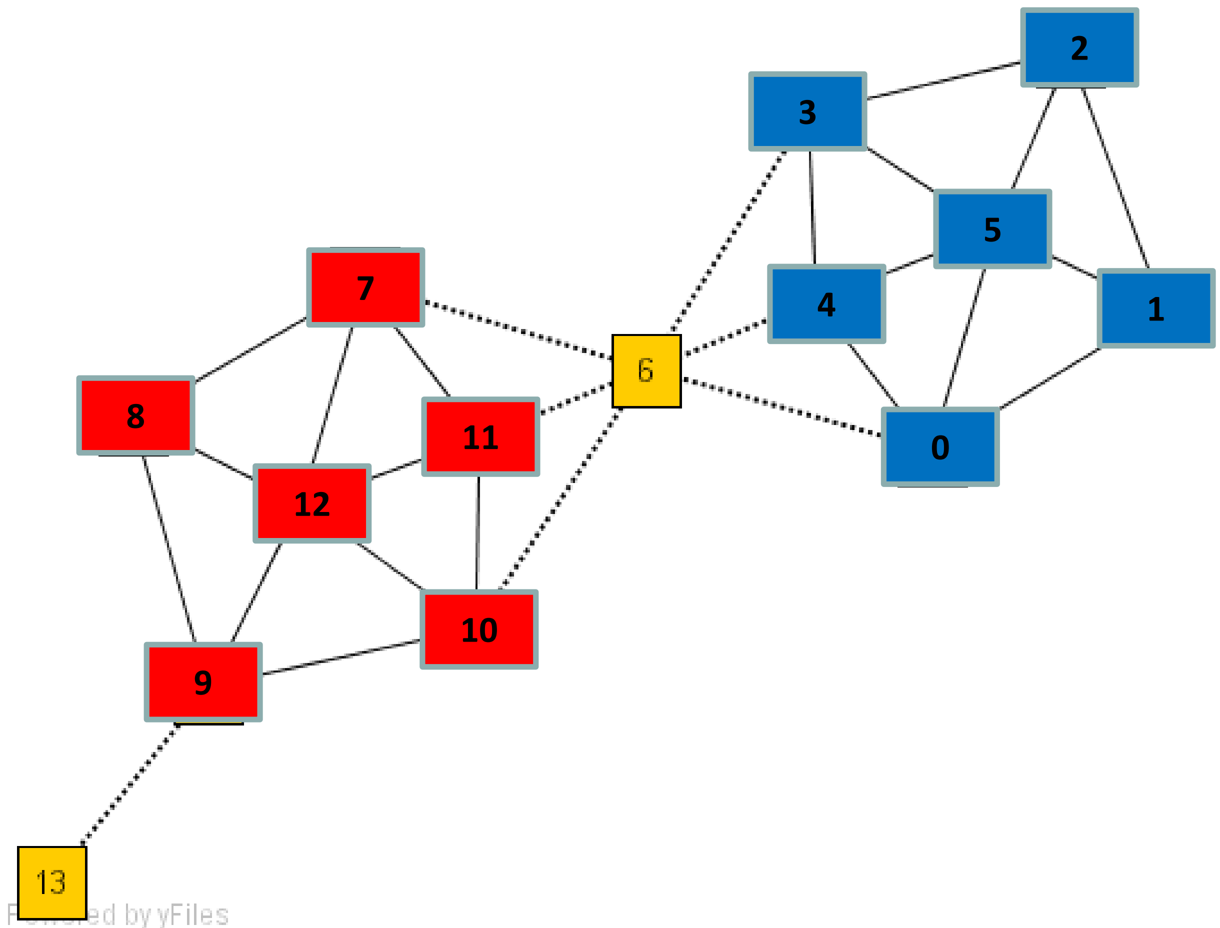
The Label Propagation Algorithm

- Step 1: Propagate $Y \leftarrow TY$
  - $Y_i = \sum_j T_{ij} Y_j = \sum_j P(j \rightarrow i) Y_j$
- Step 2: Row-normalize Y
  - The summation of the probability of each object belonging to each class is 1
- Step 3: Reset the labels for the labeled nodes. Repeat 1-3 until Y converges

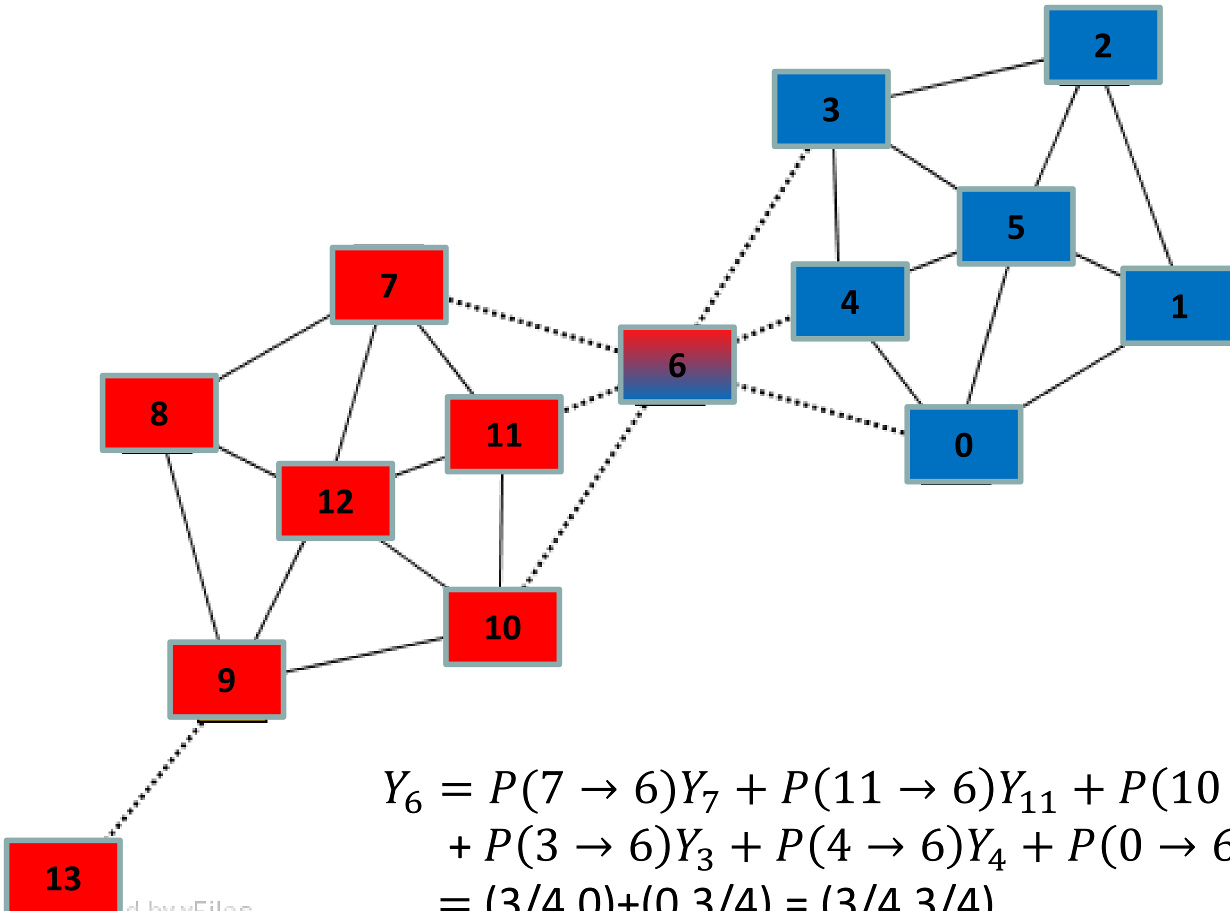
Example: Iter = 0



Example: Iter = 1



Example: Iter = 2



$$Y_6 = P(7 \to 6)Y_7 + P(11 \to 6)Y_{11} + P(10 \to 6)Y_{10}$$
$$+ P(3 \to 6)Y_3 + P(4 \to 6)Y_4 + P(0 \to 6)Y_0$$
$$= (3/4,0)+(0,3/4) = (3/4,3/4)$$

After normalization, $Y_6 = (\frac{1}{2}, \frac{1}{2})$

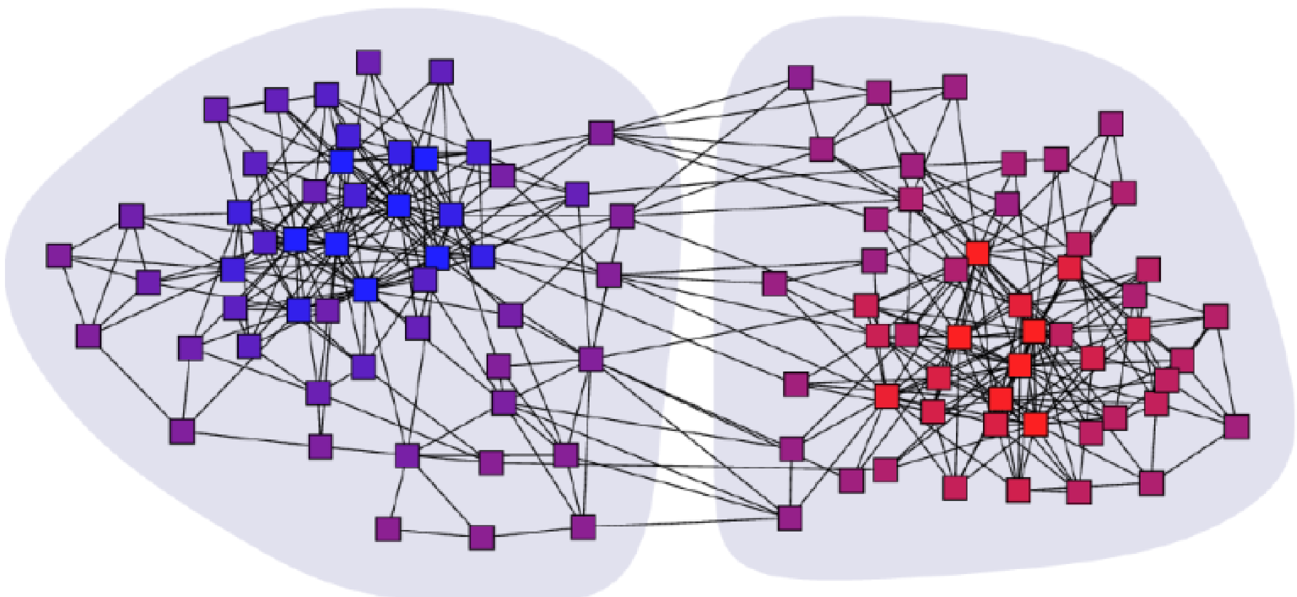
- Repeat until converge…

Mining Graph/Network Data

- Introduction to Graph/Network Data
- PageRank
- Classification via Label Propagation
- Spectral Clustering
- Summary

Clustering Graphs and Network Data

- Applications
  - Bi-partite graphs, e.g., customers and products, authors and conferences
  - Web search engines, e.g., click through graphs and Web graphs
  - Social networks, friendship/coauthor graphs



Clustering books about politics [Newman, 2006]

Spectral Clustering

- Reference: ICDM'09 Tutorial by Chris Ding
- Example:
  - Clustering supreme court justices according to

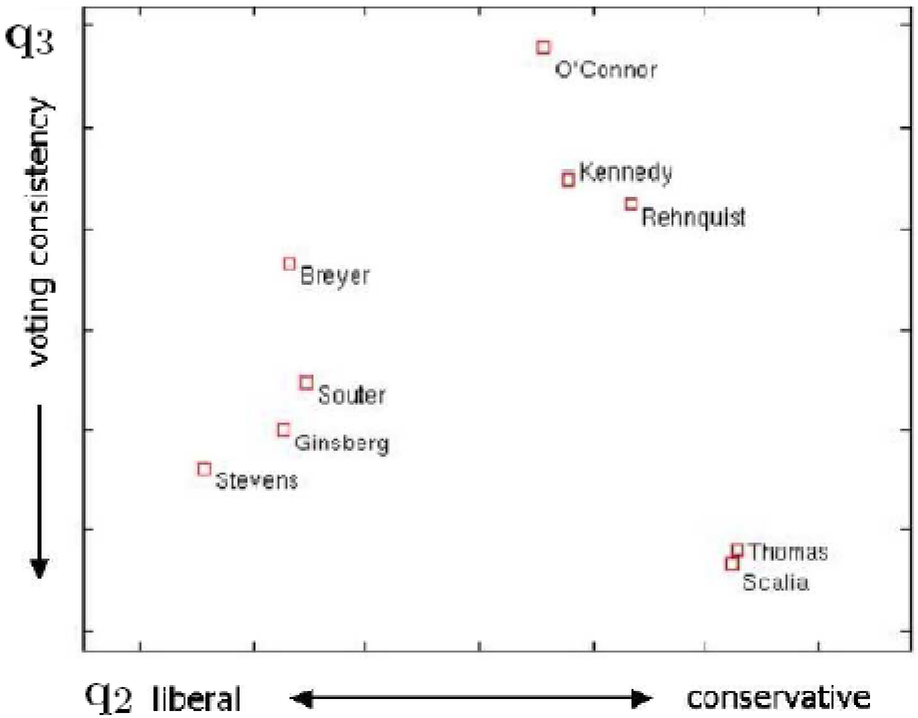
W =

Number of times (%) two Justices voted in agreement

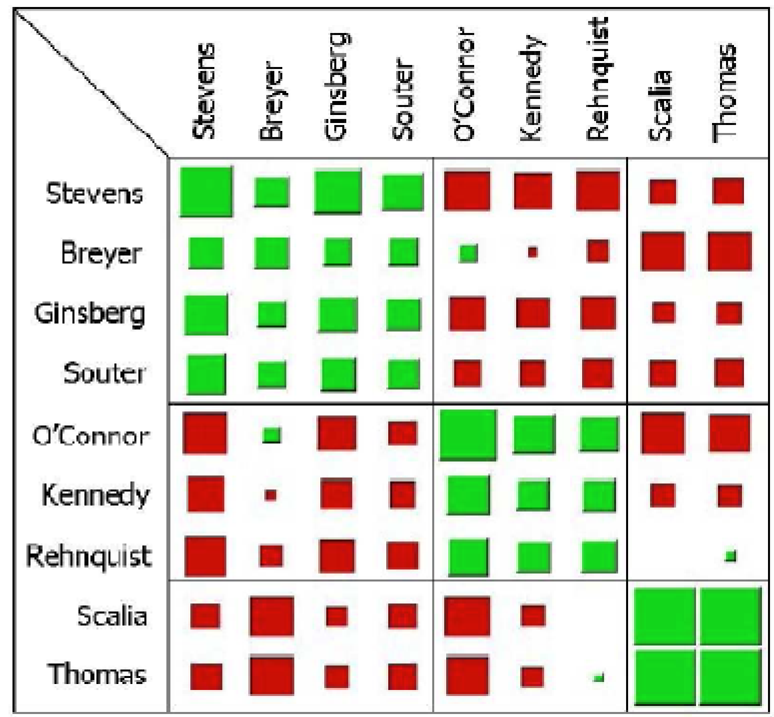
	Ste	Bre	Gin	Sou	O'Co	Ken	Reh	Sca	Tho
Stevens	–	62	66	63	33	36	25	14	15
Breyer	62	–	72	71	55	47	43	25	24
Ginsberg	66	72	–	78	47	49	43	28	26
Souter	63	71	78	–	55	50	44	31	29
O'Connor	33	55	47	55	–	67	71	54	54
Kennedy	36	47	49	50	67	–	77	58	59
Rehnquist	25	43	43	44	71	77	–	66	68
Scalia	14	25	28	31	54	58	66	–	79
Thomas	15	24	26	29	54	59	68	79	–

Table 1: From the voting record of Justices 1995 Term – 2004 Term, the number of times two justices voted in agreement (in percentage). (Data source: from July 2, 2005 *New York Times*. Originally from *Legal Affairs*; *Harvard Law Review*)

Example: Continue



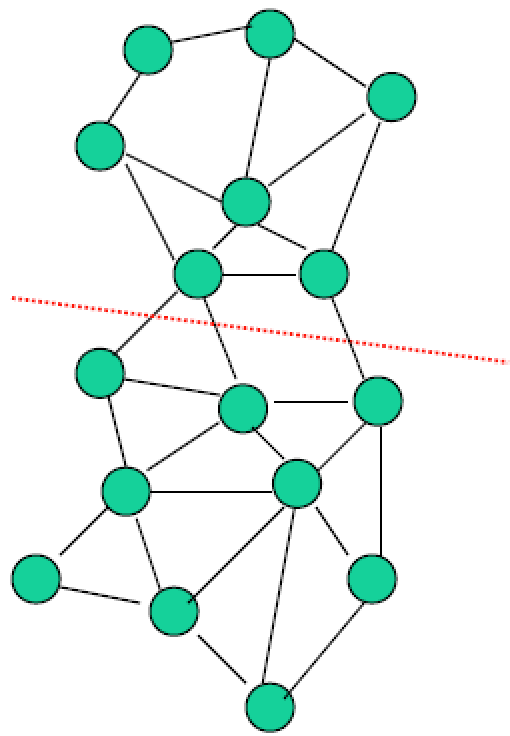
$$C = q_2 q_2^T + q_3 q_3^T$$



- Three groups in the Supreme Court:
  - Left leaning group, center-right group, right leaning group.

Spectral Graph Partition

- Min-Cut
  - Minimize the # of cut of edges



Objective Function

2-way Spectral Graph Partitioning

Partition membership indicator: $q_i = \begin{cases} 1 & \textbf{if } i \in A \\ -1 & \textbf{if } i \in B \end{cases}$

$$J = CutSize = \frac{1}{4}\sum_{i,j} w_{ij}[q_i - q_j]^2$$

$$= \frac{1}{4}\sum_{i,j} w_{ij}[q_i^2 + q_j^2 - 2q_i q_j] = \frac{1}{2}\sum_{i,j} q_i[d_i\delta_{ij} - w_{ij}]q_j$$

$$= \frac{1}{2} q^T (D - W) q$$

Relax indicators q_i from discrete values to continuous values, the solution for min $J(q)$ is given by the eigenvectors of

$$(D - W)q = \lambda q$$

(Fiedler, 1973, 1975)

(Pothen, Simon, Liou, 1990)

Algorithm

- Step 1:
  - Calculate Graph Laplacian matrix: $L = D - W$
- Step 2:
  - Calculate the second eigenvector q
- Step 3:
  - Bisect q (e.g., 0) to get two clusters

$$(D - W)q = \lambda q$$

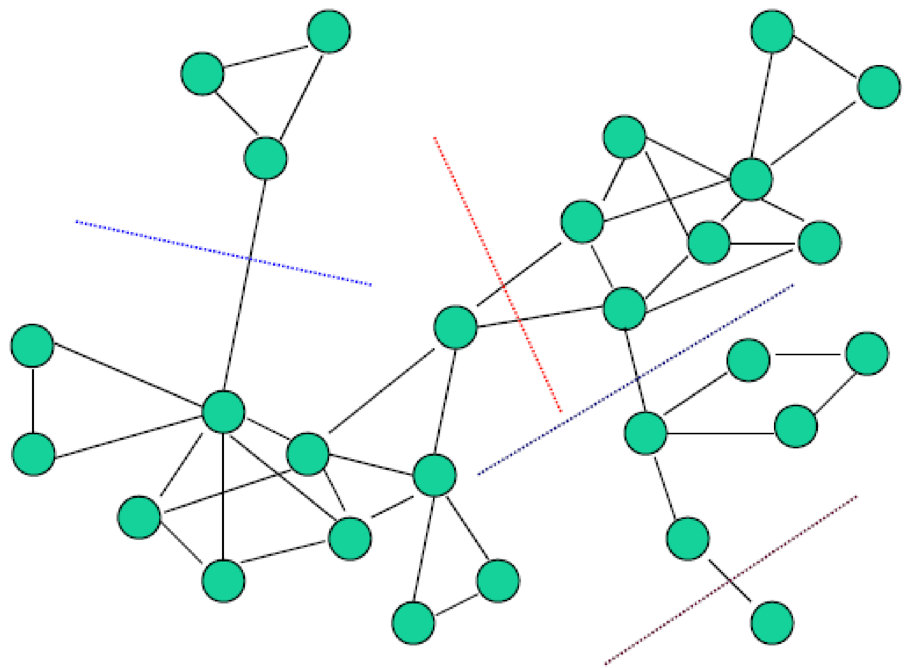
(Fiedler, 1973, 1975)

(Pothen, Simon, Liou, 1990)

*Minimum Cut with Constraints

minimize cutsize without explicit size constraints

But where to cut ?



Need to balance sizes

*New Objective Functions

- Ratio Cut (Hangen & Kahng, 1992)

$$s(A,B) = \sum_{i \in A} \sum_{j \in B} w_{ij}$$

$$J_{Rcut}(A,B) = \frac{s(A,B)}{|A|} + \frac{s(A,B)}{|B|}$$

- Normalized Cut (Shi & Malik, 2000)

$$d_A = \sum_{i \in A} d_i$$

$$J_{Ncut}(A,B) = \frac{s(A,B)}{d_A} + \frac{s(A,B)}{d_B}$$

$$= \frac{s(A,B)}{s(A,A)+s(A,B)} + \frac{s(A,B)}{s(B,B)+s(A,B)}$$

- Min-Max-Cut (Ding et al, 2001)

$$J_{MMC}(A,B) = \frac{s(A,B)}{s(A,A)} + \frac{s(A,B)}{s(B,B)}$$

Other References

- A Tutorial on Spectral Clustering by U. Luxburg http://www.kyb.mpg.de/fileadmin/user_upload/files/publications/attachments/Luxburg07_tutorial_4488%5B0%5D.pdf

Mining Graph/Network Data

- Introduction to Graph/Network Data
- PageRank
- Classification via Label Propagation
- Spectral Clustering
- Summary

Summary

- Ranking on Graph / Network
  - PageRank
- Classification via label propagation
- Clustering
  - Spectral clustering

Announcements

- Presentation next week
  - Team 1-4: Monday
  - Team 5-8: Wednesday

- Project report, code, and data
  - 6/12

- Teaching evaluation form