



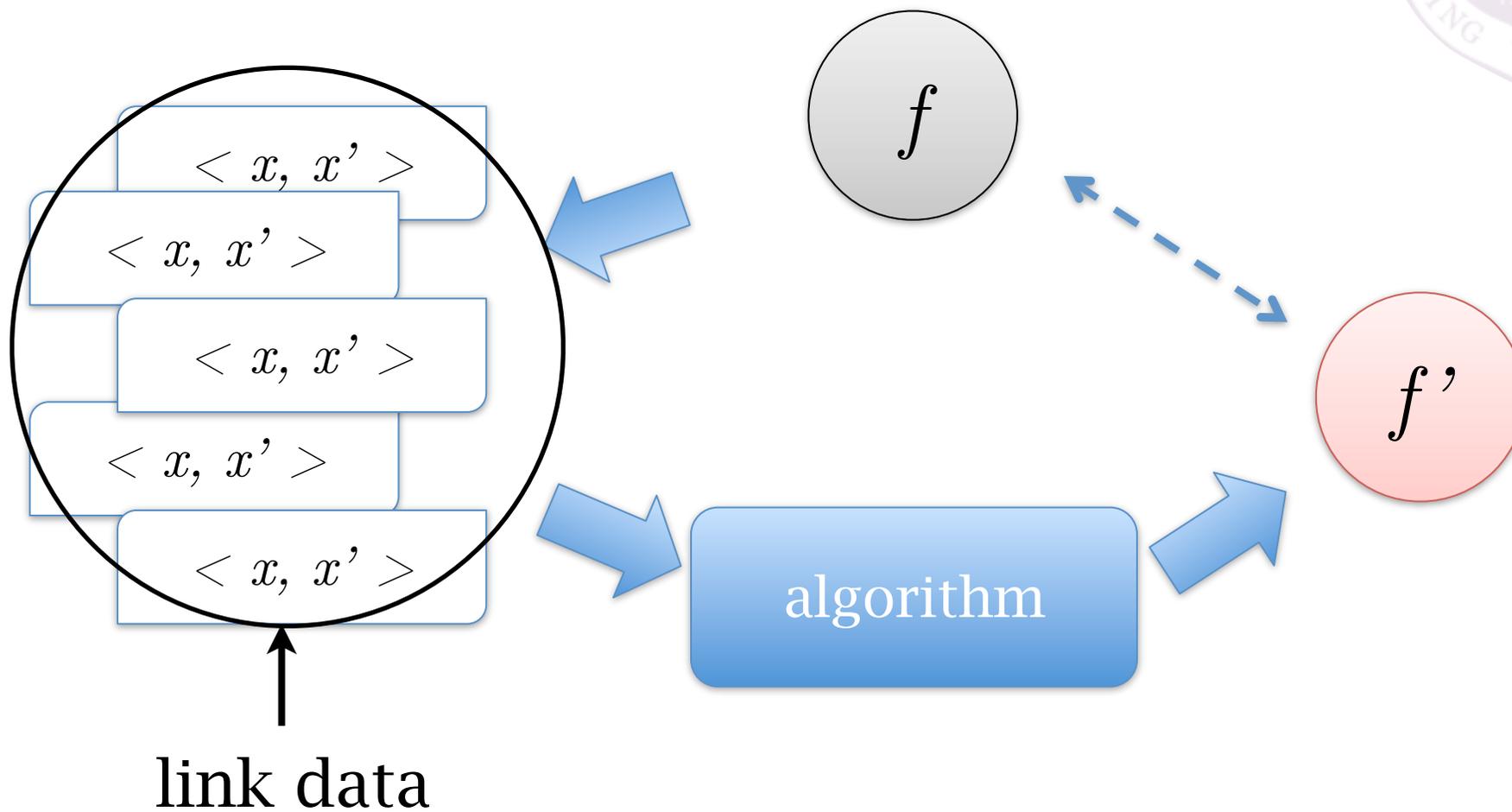
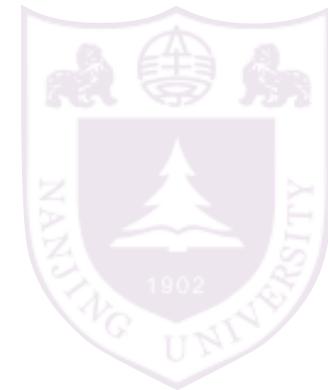
Lecture 13: Data Mining VI

Mining Link Data

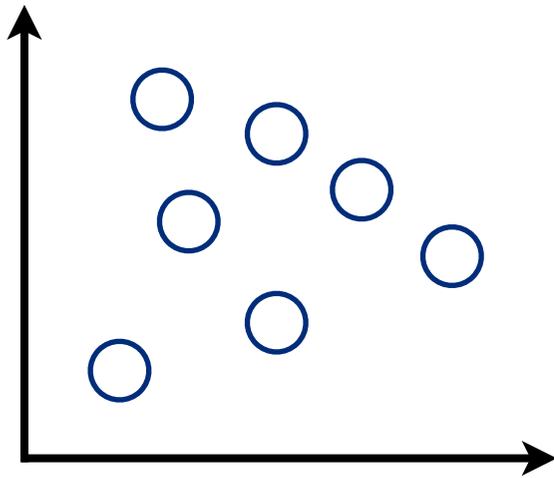
http://cs.nju.edu.cn/yuy/course_dm14ms.ashx



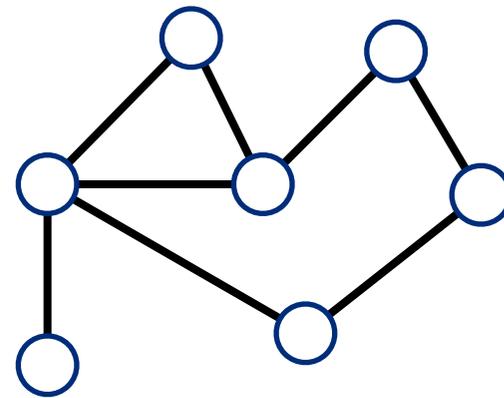
Position



What is link data



vector data



link data
= graph

chain
tree
acyclic graph
graph
multi-graph
...
directed
undirected

nodes may have features, but we focus on
the information of the edges at the moment

Why care links



pervasive and easy to obtain



friendship

hyperlink



any relationship...

Blah blah blah blah blah, blah blah blah blah blah blah. Blah blah blah blah blah blah blah. Blah blah blah, blah blah blah blah. According to Lee (2005), something very interesting was the result. Something something something, something something. Blah blah blah blah. Smith (2005) reports on some key effects of e-something on something, and suggests another interesting point. Something something blah something.

However a recent study indicates something even more interesting: blah something blah something blah something (Jones *et al*, 2006). Blah blah, blah blah, blah blah.

Reference List

Jones, C., Smith, A., Garcia, D. & Lee, A. B. (2006). Challenges in e-something. *Something Interesting*, 40, pp50-55.

Lee, A. B. (2005). *An Organisational Theory Of Something*. New York, NY: Reference Books.

Smith, A. (2005). E-something. In: Black, A. & White, B. (Eds.), *An Introduction To Something*, 30-52. Edinburgh: Textbook.

citation

Why care links

more explicit semantic



friendship

(city, job, age, salary)

are they friends?

sometimes feature vectors are used to obtain links

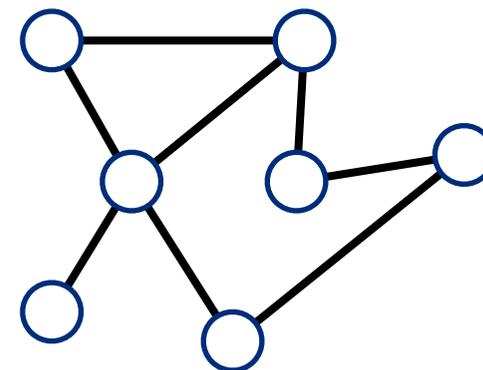
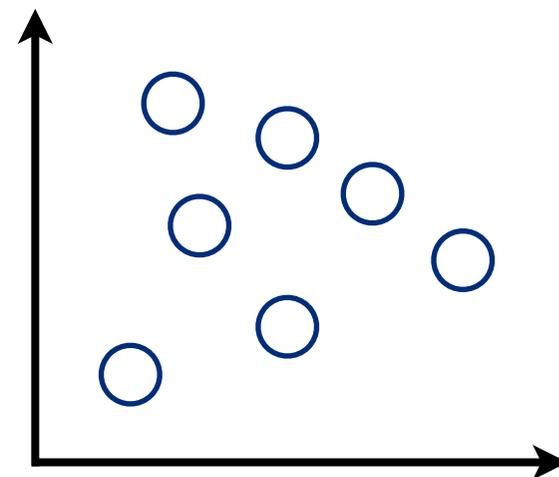
e.g. find neighbor instances

Why care links

relax i.i.d. assumption

in supervised learning, we commonly assume objects are i.i.d. drawn from a fixed distribution

link data explicitly expresses the relationship among objects



Goals in mining link data

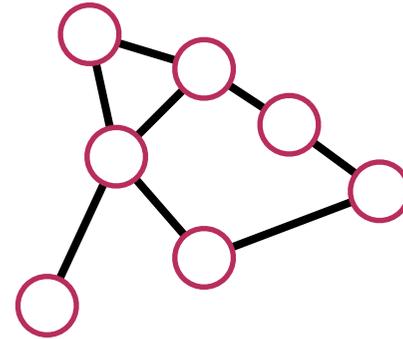


many tasks could be performed with link data

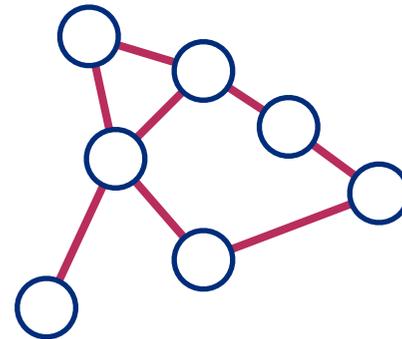
object ranking

object classification

object clustering



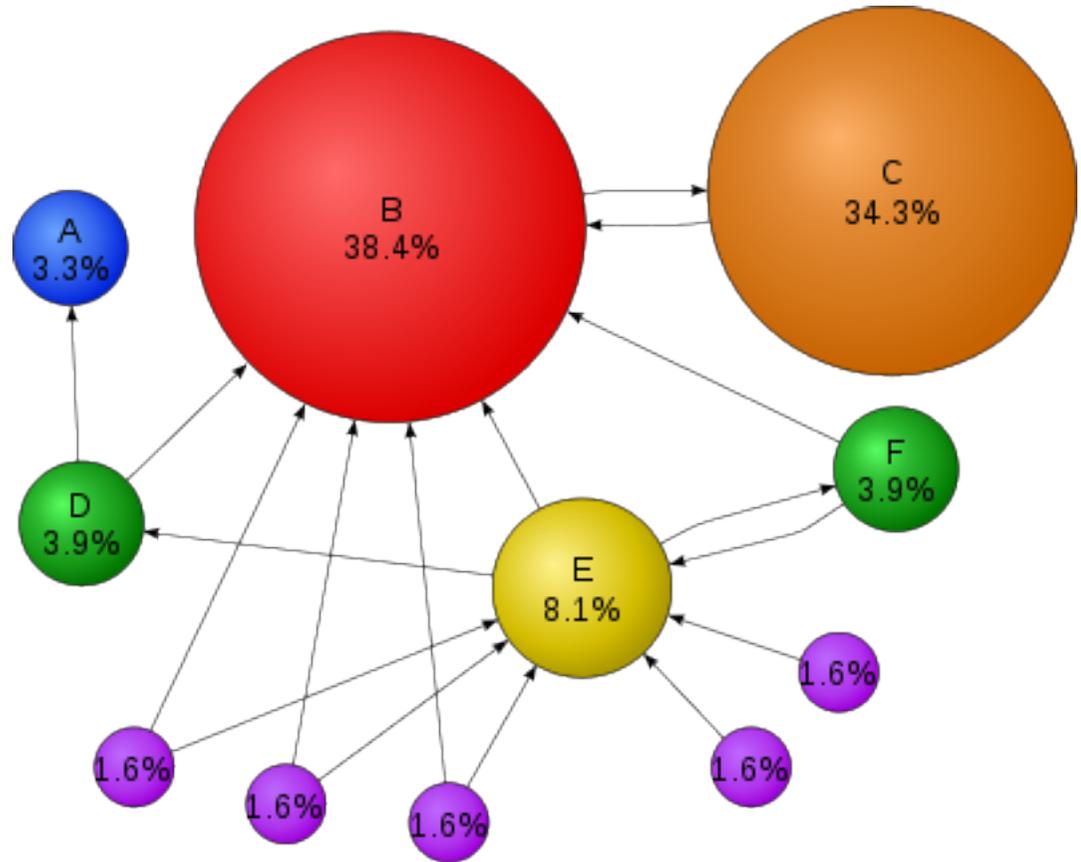
link prediction



Object ranking



ranking the importance of nodes in a directed graph



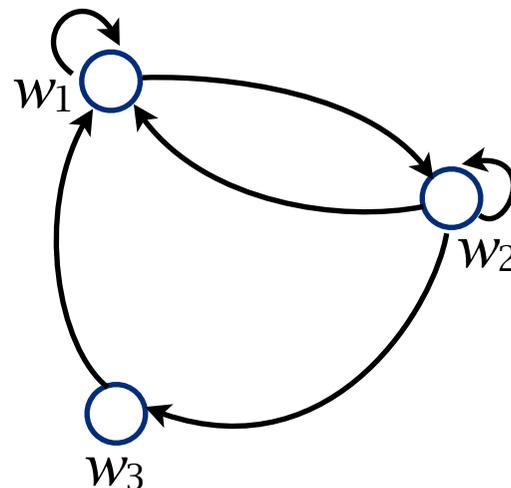
Object ranking



PageRank [PagePage, et al., 1998]

Randomly surf in the web

The importance of a web be the fraction of time staying in the web after infinite surfing time



transition matrix M

	w ₁	w ₂	w ₃
w ₁	0.5	0.5	0
w ₂	0.33	0.33	0.33
w ₃	1	0	0

current state w_1 , next state: $(1,0,0) * M = (0.5, 0.5, 0)$

next state: $(0.5, 0.5, 0) * M = (1, 0, 0) * M * M = (0.416, 0.416, 0.167)$

next state: $(1, 0, 0) * M^3 = (0.514, 0.347, 0.139)$

after 10 steps: $(0.5, 0.375, 0.125)$ stationary distribution

Object ranking

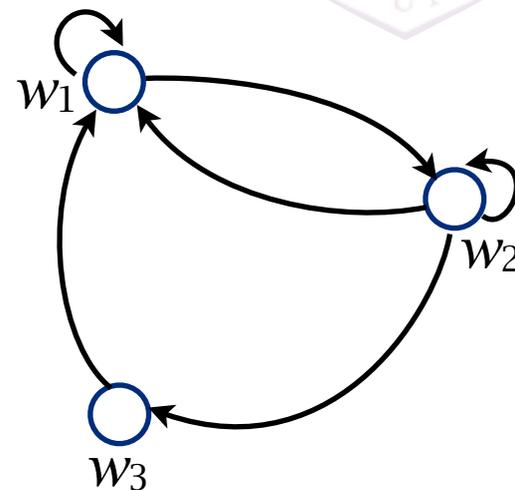


PageRank [Page, et al., 1998]

Let \mathbf{r} be the stationary distribution:

$$\mathbf{r} = M^T \mathbf{r}$$

\mathbf{r} is the eigenvector of M^T with the eigenvalue 1



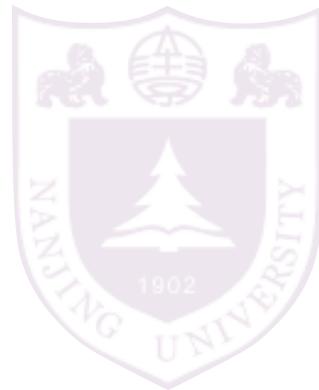
transition matrix M

	w1	w2	w3
w1	0.5	0.5	0
w2	0.33	0.33	0.33
w3	1	0	0

A PageRank voting view:

$$\mathbf{r}(x_i) = \mathbf{r}(x_1)P(x_i|x_1) + \dots + \mathbf{r}(x_n)P(x_i|x_n)$$

Object ranking



PageRank [Page, et al., 1998]

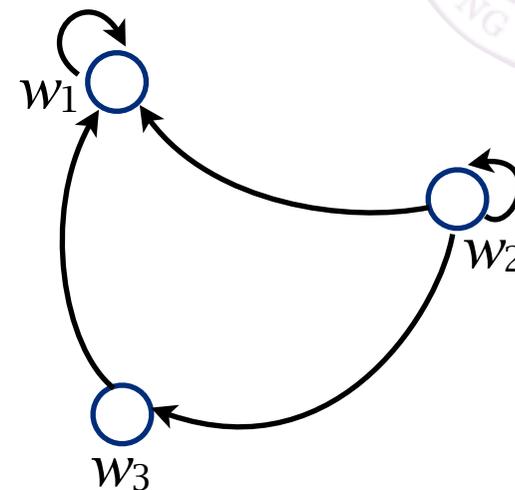
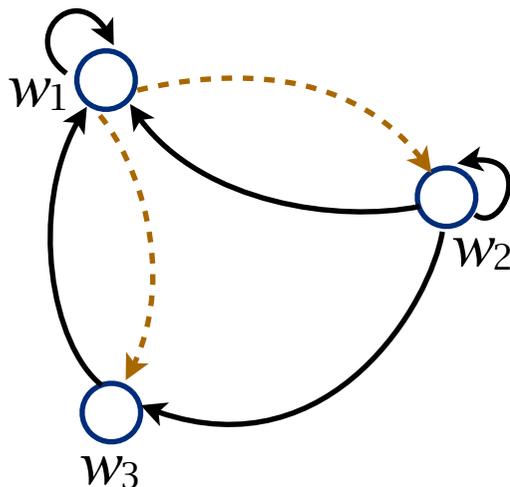
The problem with absorbing states

PageRank:

$$w_1 = 1, w_2 = w_3 = 0$$

Add a full graph:

jump to a random state with a small probability (restart)



transition matrix M

	w_1	w_2	w_3
w_1	1	0	0
w_2	0.33	0.33	0.33
w_3	1	0	0

Object ranking



PageRank [Page, et al., 1998]

Damping factor: the surfing process restarts with probability $1-d$ ($d=0.85$)

A PageRank voting view:

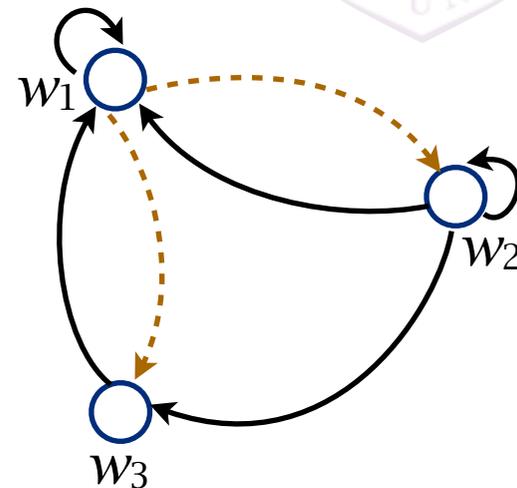
$$\mathbf{r}(x_i) = (1-d)\frac{1}{n} + d(\mathbf{r}(x_1)P(x_i|x_1) + \dots + \mathbf{r}(x_n)P(x_i|x_n))$$

Matrix form:

$$\mathbf{r} = \frac{1-d}{n}\mathbf{1} + dM^T\mathbf{r}$$

\mathbf{r} solution: $\mathbf{r} = (I - dM^T)^{-1}\frac{1-d}{n}\mathbf{1}$

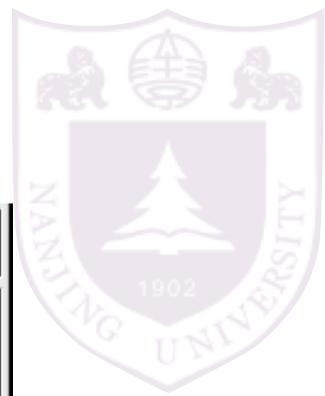
recursive solution: $\mathbf{r}_{t+1} = \frac{1-d}{n}\mathbf{1} + dM^T\mathbf{r}_t$



transition matrix M

	w_1	w_2	w_3
w_1	0.5	0.5	0
w_2	0.33	0.33	0.33
w_3	1	0	0

Object ranking



Multi Search university [Next!](#) [\[national parks\]](#)

10 results clustering on Search

Query: **university**
11 Results Returned
Showing Results From 0 to 10

Stanford University Homepage
74.79% <http://www.stanford.edu>
4k - 3591993 - 010397

Stanford University: Portfolio Collection
65.78% <http://www.stanford.edu/home/administration/portfolio.html>
3k - 3591993 - 010397

University of Illinois at Urbana-Champaign
73.26% <http://www.uiuc.edu>
13k - 1330195 - 010397

Indiana University
68.38% <http://www.indiana.edu>
1k - 0920195 - 010397

University of California, Irvine
68.07% <http://www.uci.edu>
3k - 1330195 - 010397

University of Minnesota
67.05% <http://www.umn.edu>
2k - 1316195 - 010397

Iowa State University Homepage
66.66% <http://www.iastate.edu>
3k - 1316195 - 010397

The University of Michigan
66.35% <http://www.umich.edu>
1k - 3591993 - 010397

Mississippi State University
66.35% <http://www.msstate.edu>
3k - 3591993 - 010397

Northwestern University: NUInfo
66.15% <http://www.nwu.edu>
3k - 1314195 - 010397

next 10

Optical Physics at the University of Oregon
Oregon Center for Optics in Science and Technology. Department of Physics, University of Oregon, Eugene OR 97403. Research Groups: Carmichael Group....
<http://optics.uoregon.edu/> - size 1K - 16 Dec 96

Carnegie Mellon University - Campus Networking
Departments. Data Communications. Data Communications is responsible for installing and maintaining all on campus networking equipment and all of...
<http://www.net.cmu.edu/> - size 4K - 19 Aug 95

Wesleyan University Computer Science Group Home Page
Computer Science Group. Wesleyan University. Welcome to the home page of the Computer Science Group at Wesleyan University. We are administratively within.
<http://www.cs.wesleyan.edu/> - size 3K - 15 Apr 96

Keio University Shonan Fujisawa Campus (SFC)
B\$\$\$N%ZIEFnF#Bt%-c%e%Q%9 (B(SFC) \$B\$N (BWWW \$B% \$BcmOU=q\$- (B \$B\$rFI\$s\$G\$/\$@5\$#\$ (B. Nihongo | English. SFC \$B>pJs (B. | \$B%a%G%#% "%,%e%?! *...
<http://www.sfc.keio.ac.jp/> - size 3K - 5 Feb 97

School of Chemistry, University of Sydney
The School of Chemistry. School of Chemistry, University of Sydney, NSW 2006 Australia International Phone: +61-2-9351-4504 Fax: +61-2-9351-3329 Australia.
<http://www.chem.su.oz.au/> - size 4K - 25 Feb 97

Mankato State University
The Campus Athletics, Campus Tour, Bookstore, Maps, Current Events... Admission & Registration Admissions, Financial Aid, Registrar's, Graduate...
<http://www.mankato.msut.edu/> - size 3K - 27 Nov 96

St. Ambrose University
Main Index: Academic Departments. Administrative Services. Campus News. Computing Services. Galvin Fine Arts Center. Internet Connections. Library...
<http://www.sau.edu/> - size 3K - 4 Feb 97

University of Washington ECSEL Projects

Figure 6: Comparison of Query for “University”

Object ranking

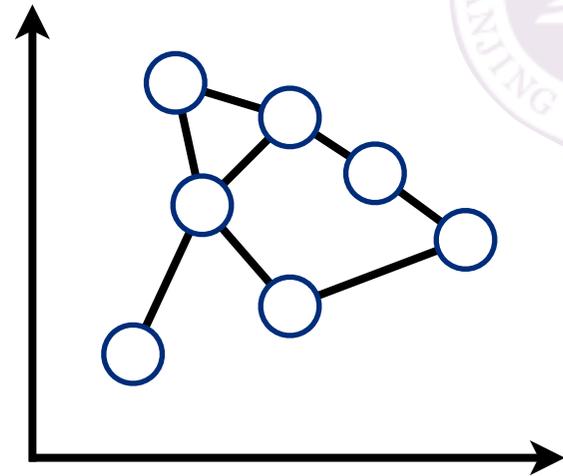


Figure 7: PageRank Proxy

[Page, et al., 1998]

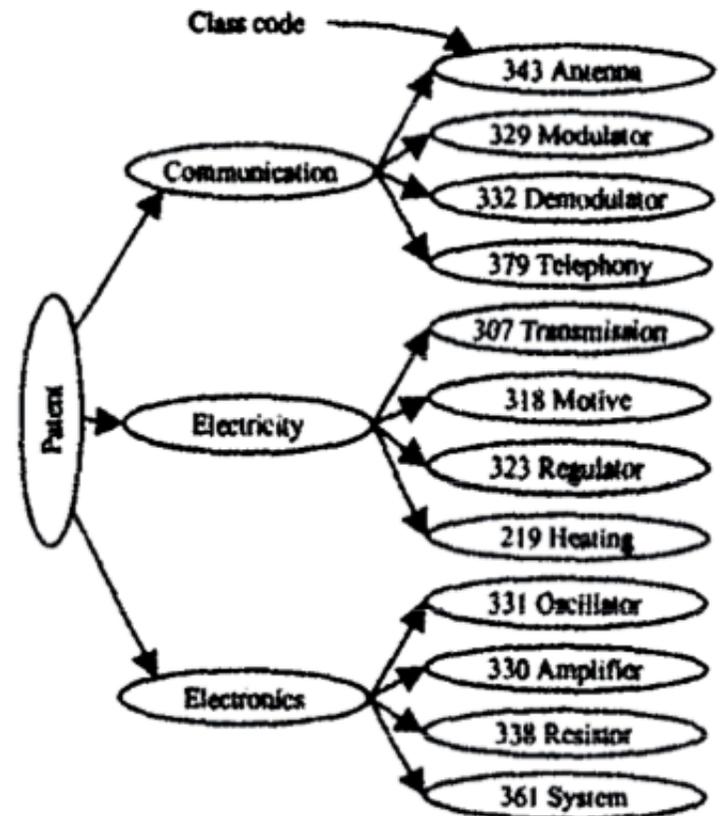
Object classification

Incorporate link information could improve the classification accuracy



Classification of web pages

[Chakrabarti, et al., SIGMOD98]



Object classification

Classification of web pages

[Chakrabarti, et al., SIGMOD98]

use pure text for classification: 36% error



Object classification

Classification of web pages

[Chakrabarti, et al., SIGMOD98]

use pure text for classification: 36% error



Object classification



Classification of web pages

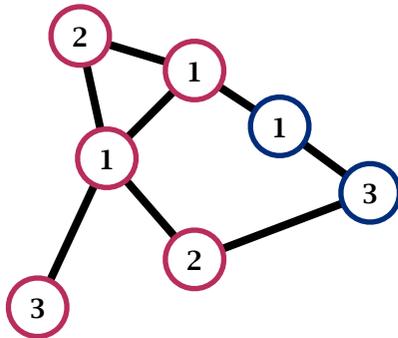
[Chakrabarti, et al., SIGMOD98]

use pure text for classification: 36% error

use neighbor predicted classes:

34% error, 22.1% error

hyperlink forms a neighborhood relationship



Given test node δ_0

Construct a radius- r subgraph $G_r(\delta_0)$ around δ_0

Assign initial classes to all $\delta \in G_r(\delta_0)$ using local text

Iterate until consistent:

Recompute the class for each δ based on
local text and class of neighbors

Object classification

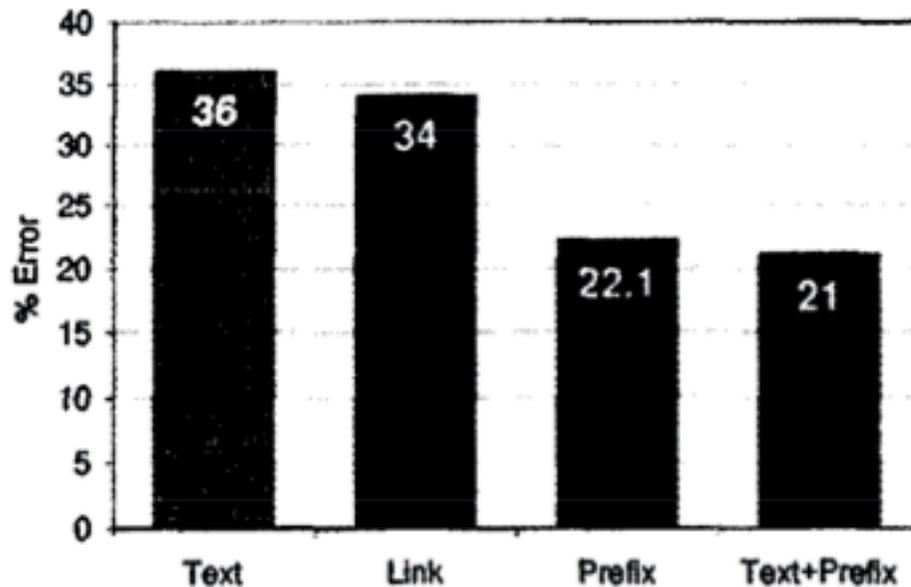


Classification of web pages

[Chakrabarti, et al., SIGMOD98]

use pure text for classification: 36% error

use neighbor predicted classes:
34% error, 22.1% error

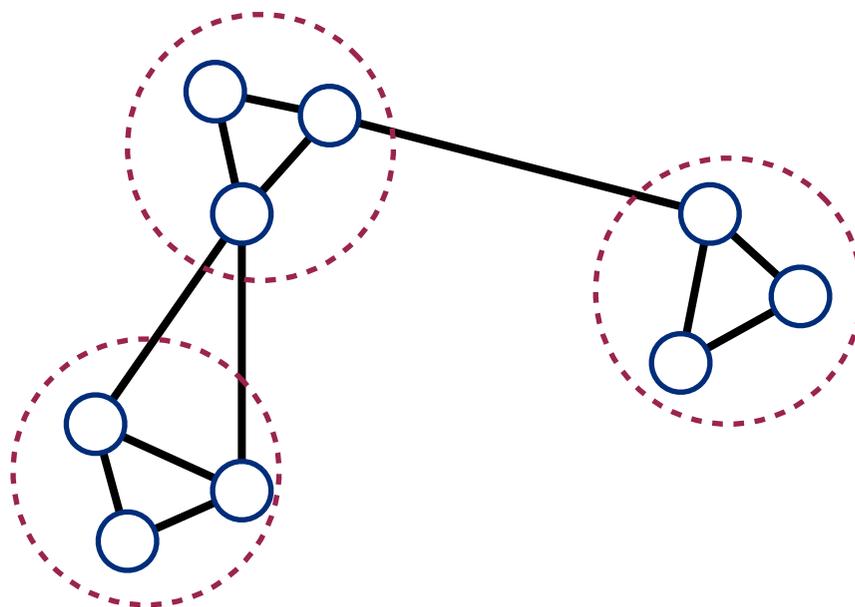


Object clustering



Clustering nodes using link information

community discovery in social networks

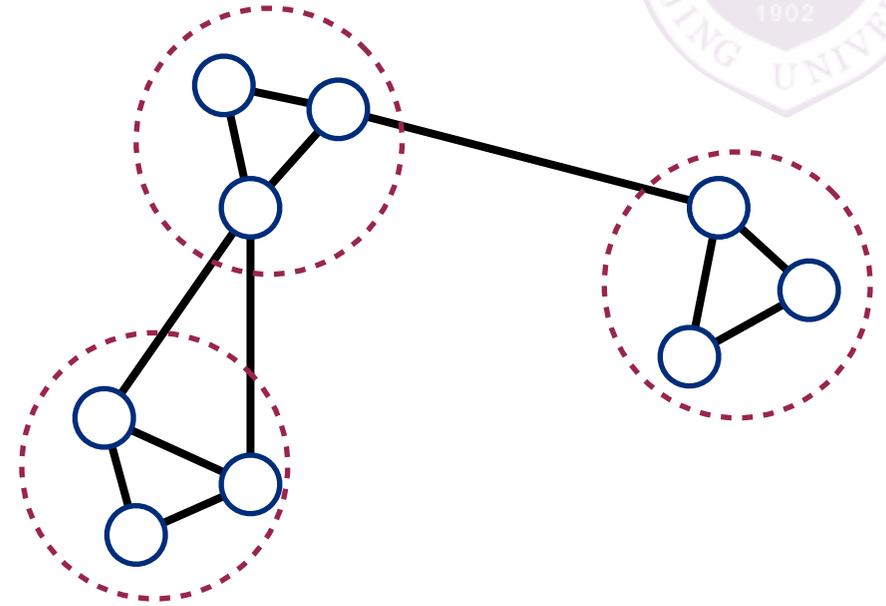


Object clustering



Presenting the graph into an adjoint matrix

1	0	1
1	1	0
0	1	1



many clustering algorithms utilize only the adjoint matrix

hierarchical clustering

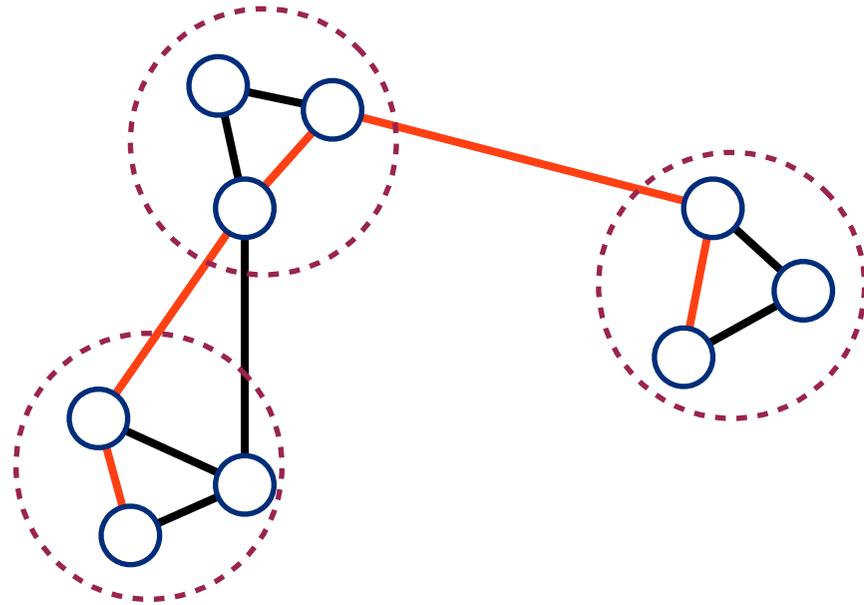
graph-cut

k-medoids

Object clustering



Defining the distance between any two nodes as the shortest path length



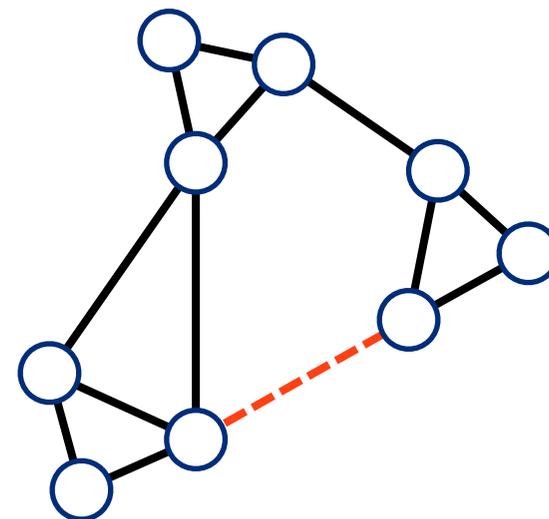
all clustering algorithms can be used

Link prediction



Predict the existence of a link between two nodes

recommendations in social network



A common solution:

compute a similarity among any pairs of nodes

the pairs with high similarity is predicted as a link

$$\text{score}(x, y)$$

Link prediction



Similarities among nodes: neighbor-based
Common neighbors [Newman, PRL'01]

$$\text{score}(x, y) := |\Gamma(x) \cap \Gamma(y)|$$

$\Gamma(x)$ is the set of neighbor nodes of x

two persons shares a lot of
friends are likely to be friends

Link prediction



Similarities among nodes: neighbor-based
Jaccard's coefficient [Salton and McGill,83]

$$\text{score}(x, y) := |\Gamma(x) \cap \Gamma(y)| / |\Gamma(x) \cup \Gamma(y)|$$

$\Gamma(x)$ is the set of neighbor nodes of x

consider the relative counting

Link prediction



Similarities among nodes: neighbor-based
Preferential attachment [Mitzenmacher, ACCCC'01]

$$\text{score}(x, y) := |\Gamma(x)| \cdot |\Gamma(y)|$$

$\Gamma(x)$ is the set of neighbor nodes of x

the probability that a new edge involves
node x is proportional to $|\Gamma(x)|$

Link prediction



Similarities among nodes: path-based

Katz [Psychometrika'53]

$$\text{score}(x, y) := \sum_{\ell=1}^{\infty} \beta^{\ell} \cdot |\text{paths}_{x,y}^{\langle \ell \rangle}|$$

$\text{paths}_{x,y}^{\langle \ell \rangle}$ is the set of all length- ℓ paths from x to y

weighted average of path length

Link prediction



Similarities among nodes: path-based
Random walk

commute time: $\text{score}(x, y) = H_{x,y} + H_{y,x}$

$H_{x,y}$ is the hitting time of random walk from x to y

normalized commute time:

$$\text{score}(x, y) := -(H_{x,y} \cdot \pi_y + H_{y,x} \cdot \pi_x)$$

π_x is the probability of x in the stationary distribution

Link prediction



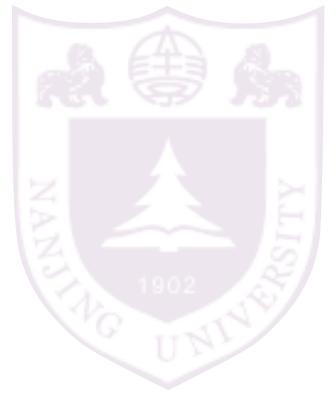
Similarities among nodes: meta methods

SimRank [Jeh and Widom, KDD02]

$$\text{similarity}(x, y) := \gamma \cdot \frac{\sum_{a \in \Gamma(x)} \sum_{b \in \Gamma(y)} \text{similarity}(a, b)}{|\Gamma(x)| \cdot |\Gamma(y)|}$$

recursively compute the similarity

习题



PageRank算法的思想是什么？