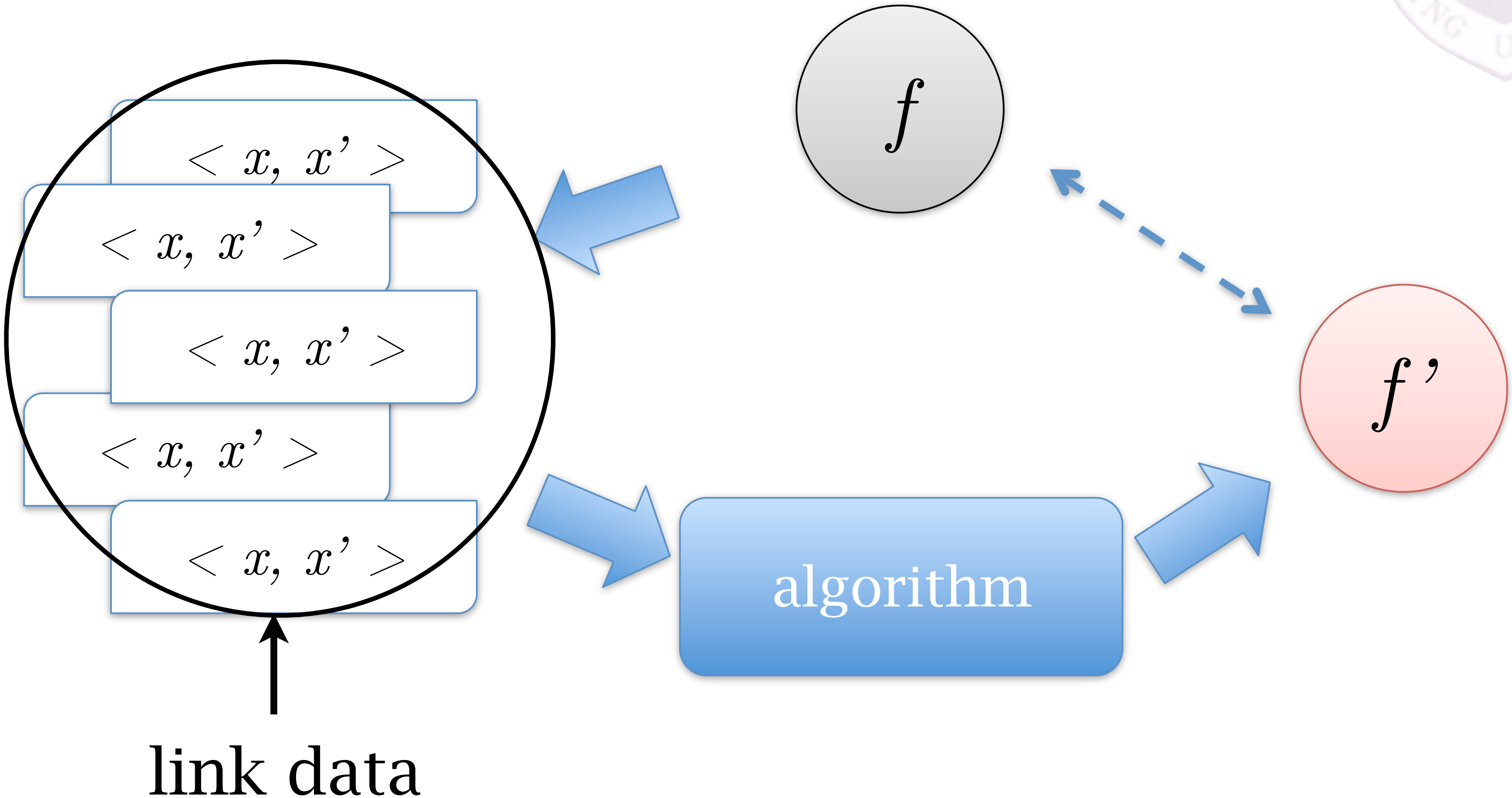
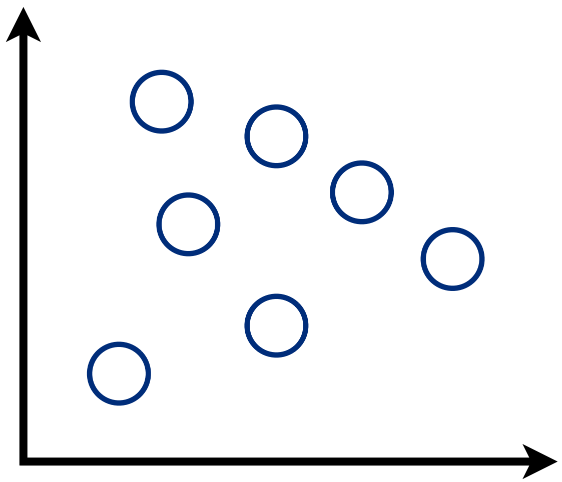
Lecture 12: Data Mining III
Mining Link Data

http://cs.nju.edu.cn/yuy/course_dm13ms.ashx

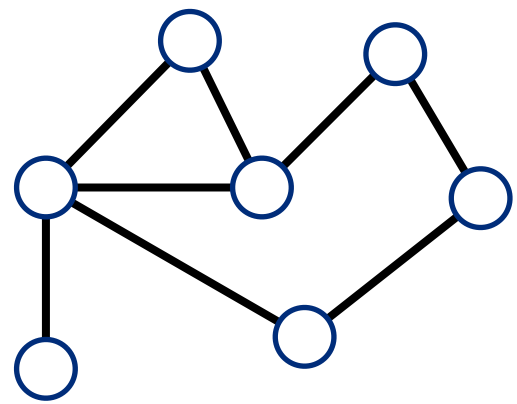
Position



What is link data



vector data



link data
= graph

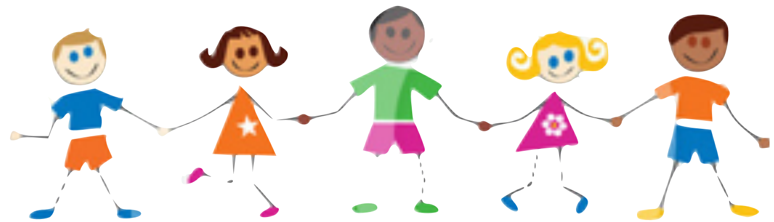
chain
tree
acyclic graph
graph
multi-graph
...
directed
undirected

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

Why care links

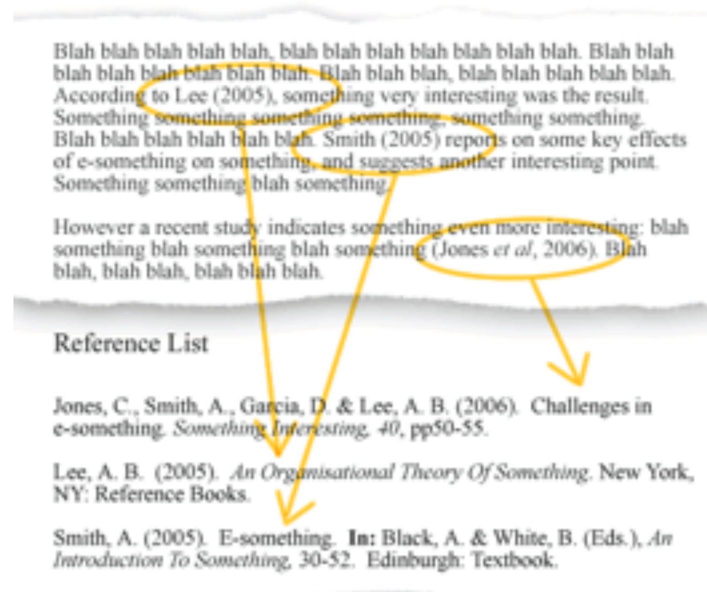
pervasive and easy to obtain

hyperlink



friendship

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 blah blah blah. According to Lee (2005), something very interesting was the result. Something something something something, something something. Blah blah 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 blah.

Reference List

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

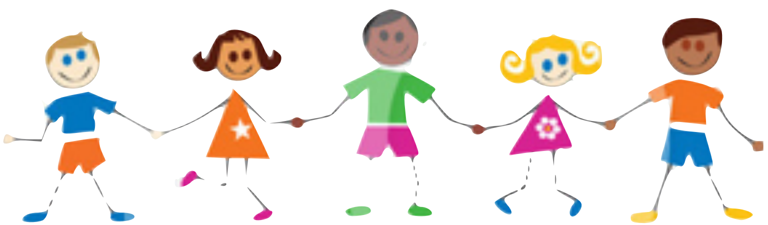
Lee, A. B. (2005). *An Organisational Theory Of Something*. New York, NY: Reference Books.

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?

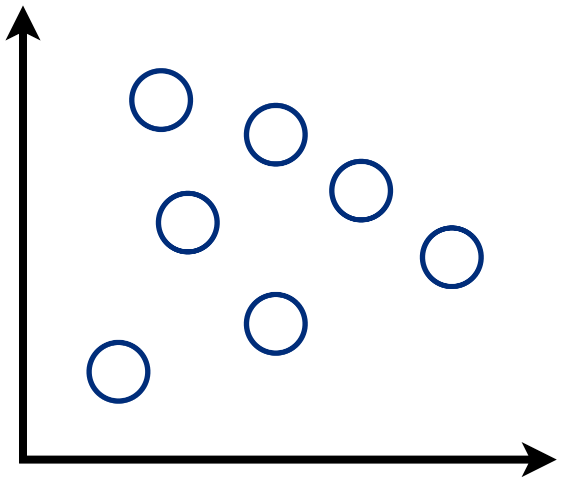
sometimes feature vectors are used to obtain links

e.g. find neighbor instances

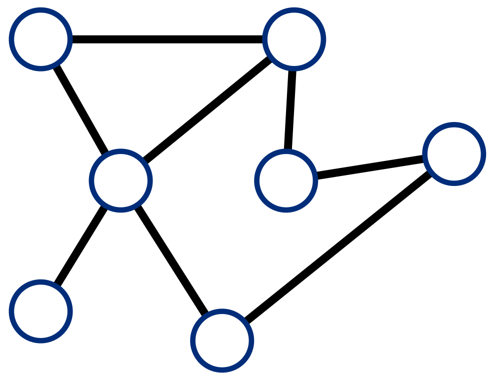
Why care links

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



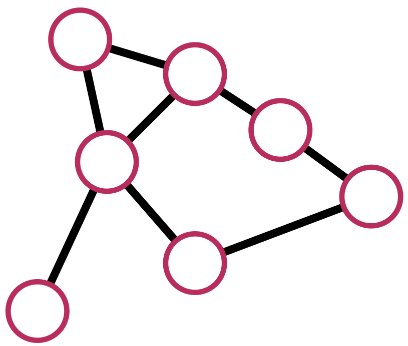
Goals in mining link data

many tasks could be performed with link data

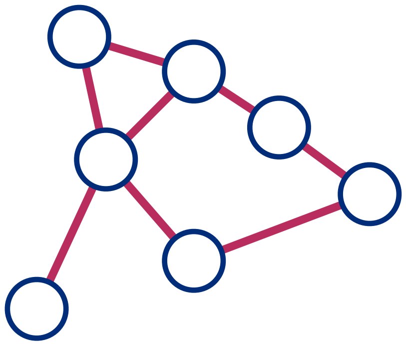
object ranking

object classification

object clustering

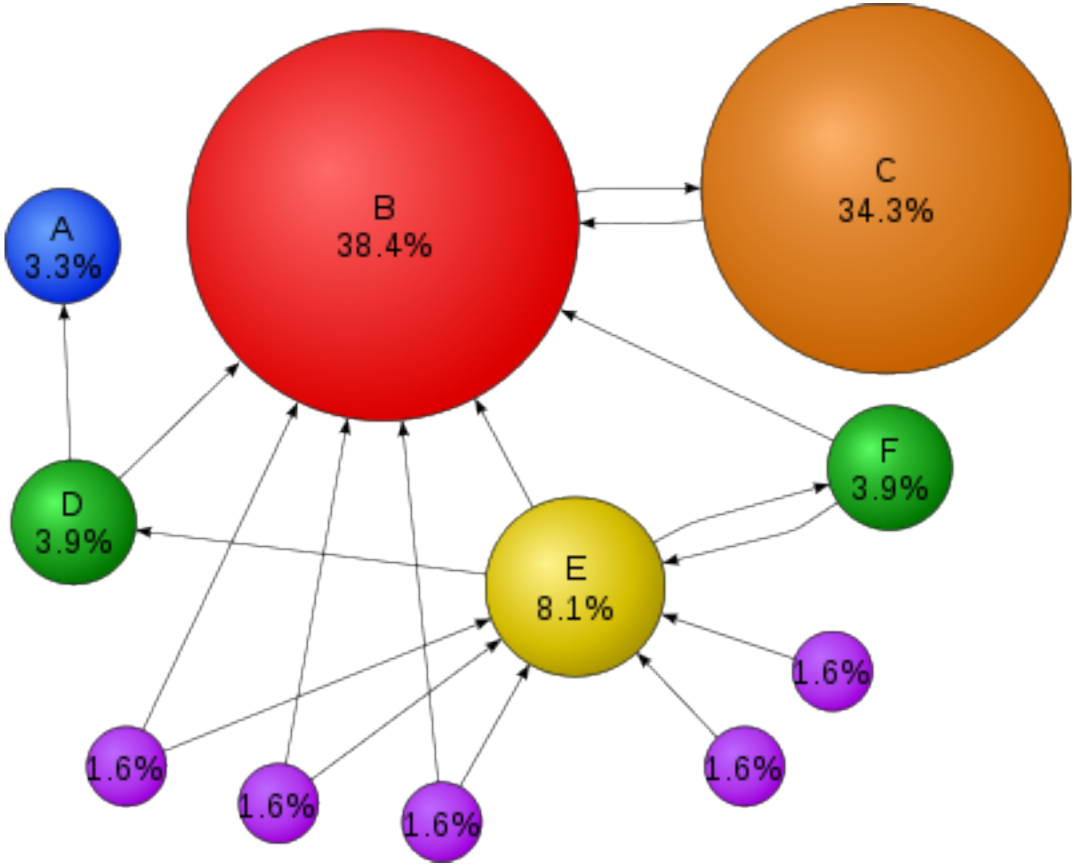


link prediction



Object ranking

ranking the importance of nodes in a directed graph

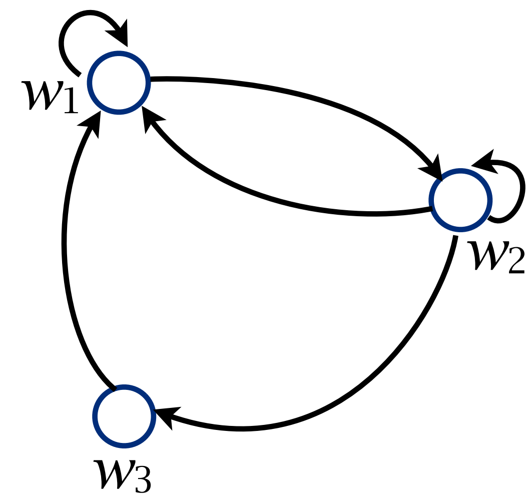


Object ranking

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

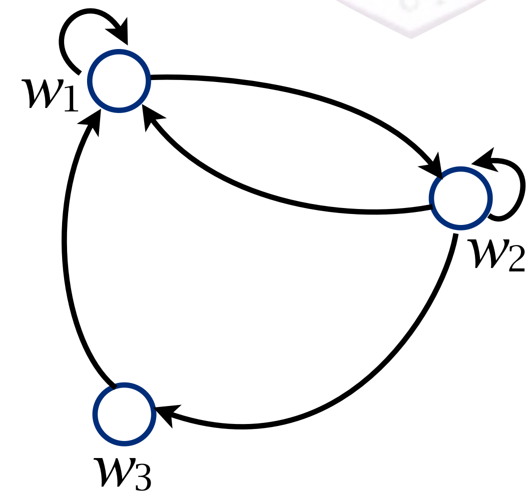
Object ranking

PageRank [Page, et al., 1998]

Let $\boldsymbol{r}$ be the stationary distribution:

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

$\boldsymbol{r}$ is the eigenvector of $M^\top$ with the eigenvalue 1



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

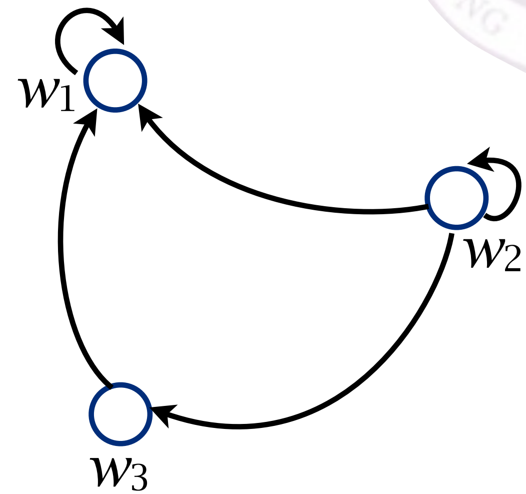
A PageRank voting view:

$$\boldsymbol{r}(x_i) = \boldsymbol{r}(x_1)P(x_i|x_1) + \ldots + \boldsymbol{r}(x_n)P(x_i|x_n)$$

Object ranking

PageRank [Page, et al., 1998]

The problem with absorbing states



PageRank:

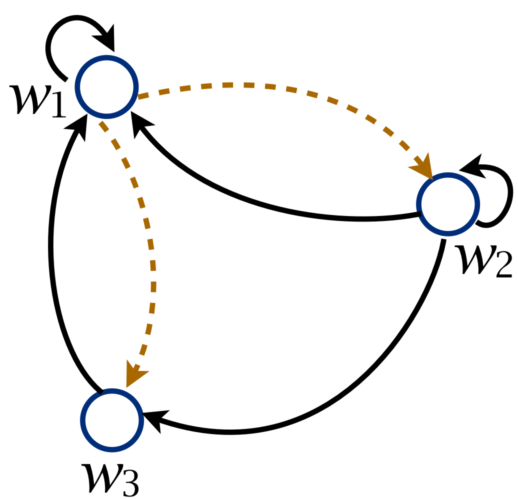
$w_1 = 1,\ w_2 = w_3 = 0$

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

Add a full graph:

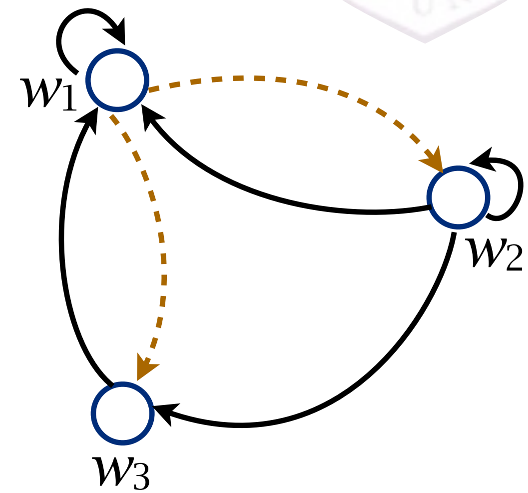
jump to a random state with a small probability (restart)



Object ranking

PageRank [Page, et al., 1998]

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



A PageRank voting view:

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

Matrix form:

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

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

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

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

Object ranking

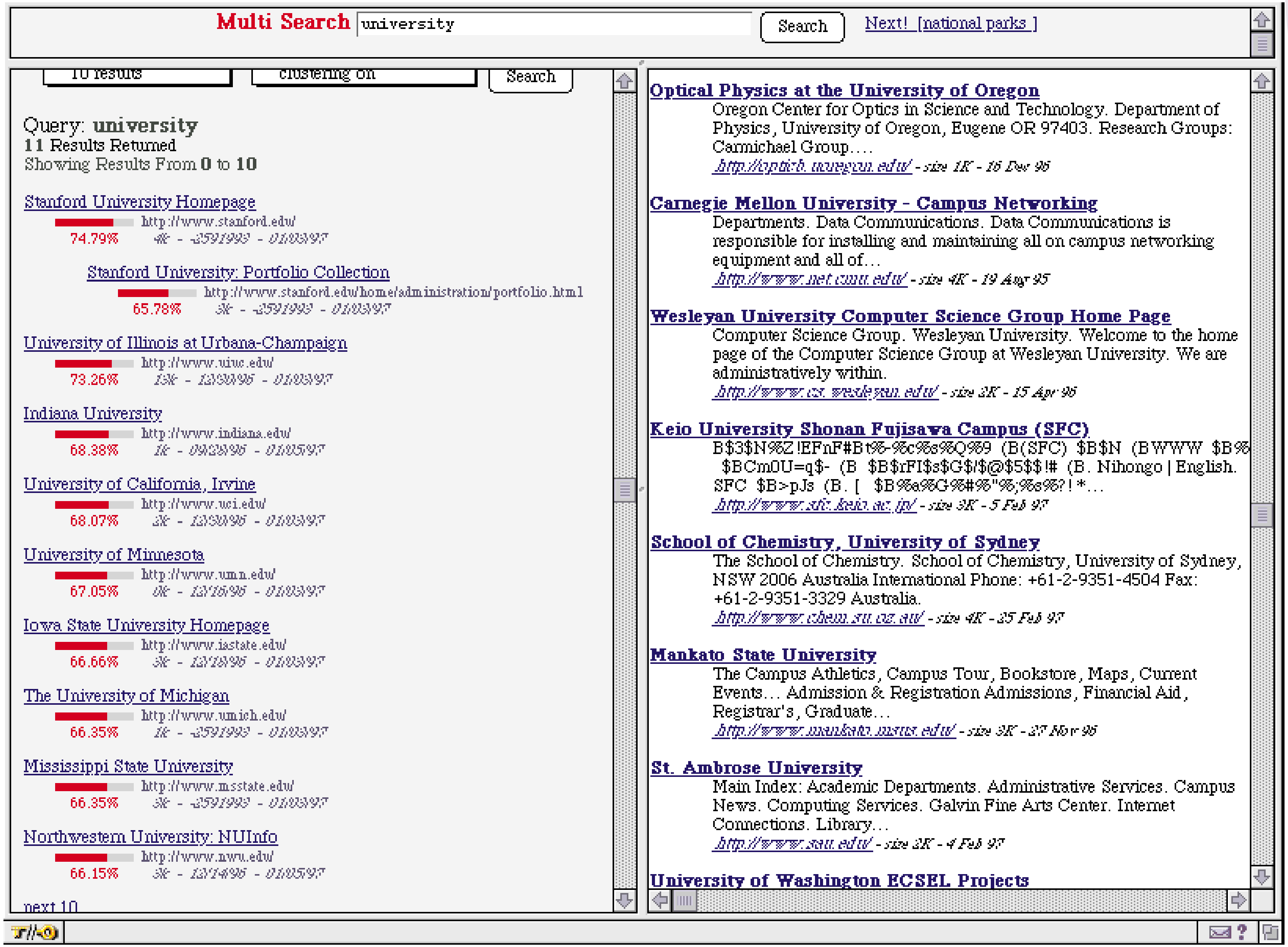
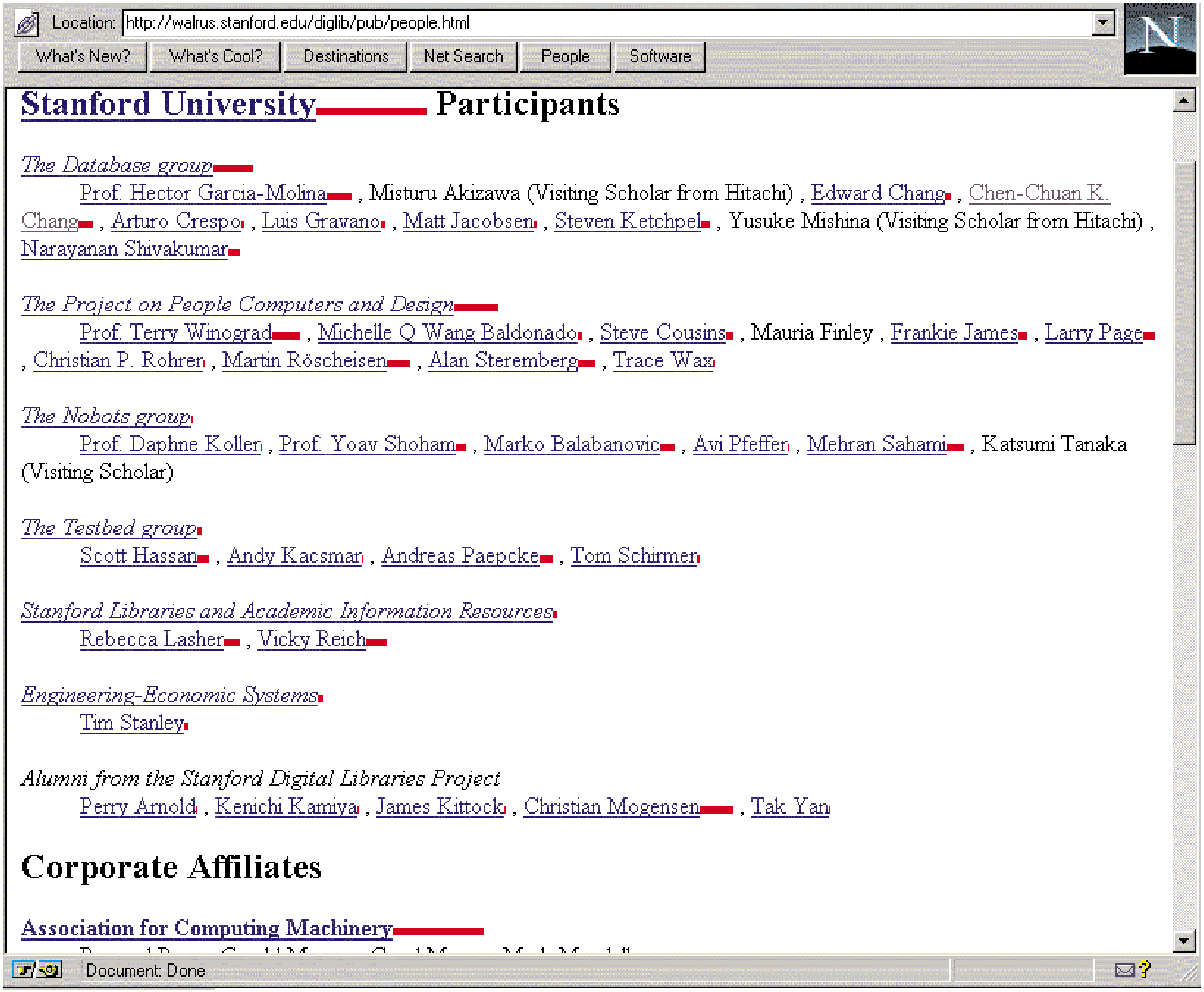


Figure 6: Comparison of Query for "University"

[Page, et al., 1998]

Object ranking



Location: http://walrus.stanford.edu/diglib/pub/people.html

What's New? | What's Cool? | Destinations | Net Search | People | Software

Stanford University Participants

The Database group

Prof. Hector Garcia-Molina , Misturu Akizawa (Visiting Scholar from Hitachi) , Edward Chang , Chen-Chuan K. Chang , Arturo Crespo , Luis Gravano , Matt Jacobsen , Steven Ketchpel , Yusuke Mishina (Visiting Scholar from Hitachi) , Narayanan Shivakumar

The Project on People Computers and Design

Prof. Terry Winograd , Michelle Q Wang Baldonado , Steve Cousins , Mauria Finley , Frankie James , Larry Page , Christian P. Rohrer , Martin Röscheisen , Alan Steremberg , Trace Wax

The Nobots group

Prof. Daphne Koller , Prof. Yoav Shoham , Marko Balabanovic , Avi Pfeffer , Mehran Sahami , Katsumi Tanaka (Visiting Scholar)

The Testbed group

Scott Hassan , Andy Kacsmar , Andreas Paepcke , Tom Schirmer

Stanford Libraries and Academic Information Resources

Rebecca Lasher , Vicky Reich

Engineering-Economic Systems

Tim Stanley

Alumni from the Stanford Digital Libraries Project

Perry Arnold , Kenichi Kamiya , James Kittock , Christian Mogensen , Tak Yan

Corporate Affiliates

Association for Computing Machinery

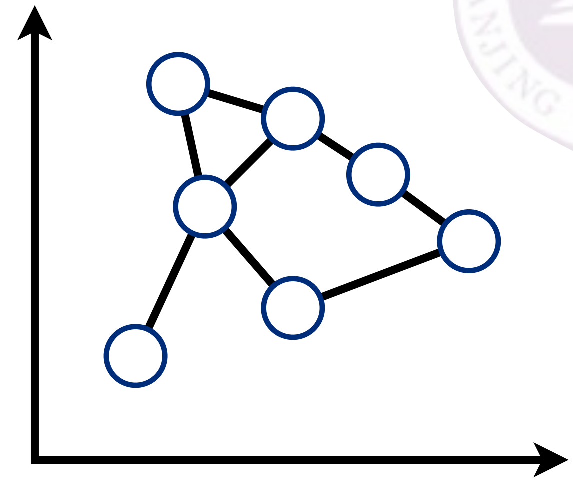
Document: Done

Figure 7: PageRank Proxy

[Page, et al., 1998]

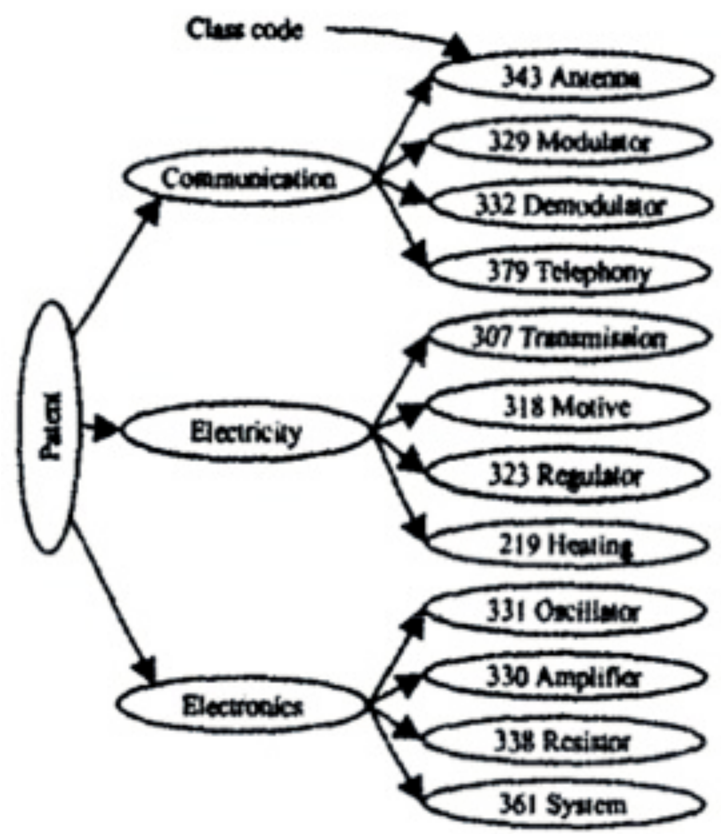
Object classification

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

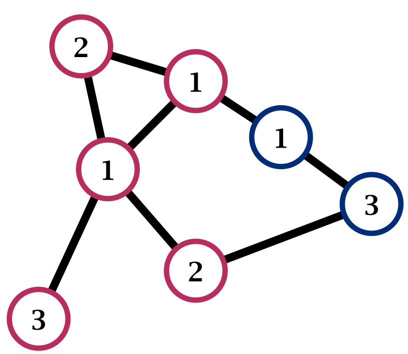
Classification of web pages

[Chakrabarti, et al., SIGMOD98]

use pure text for classification: 36% error

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

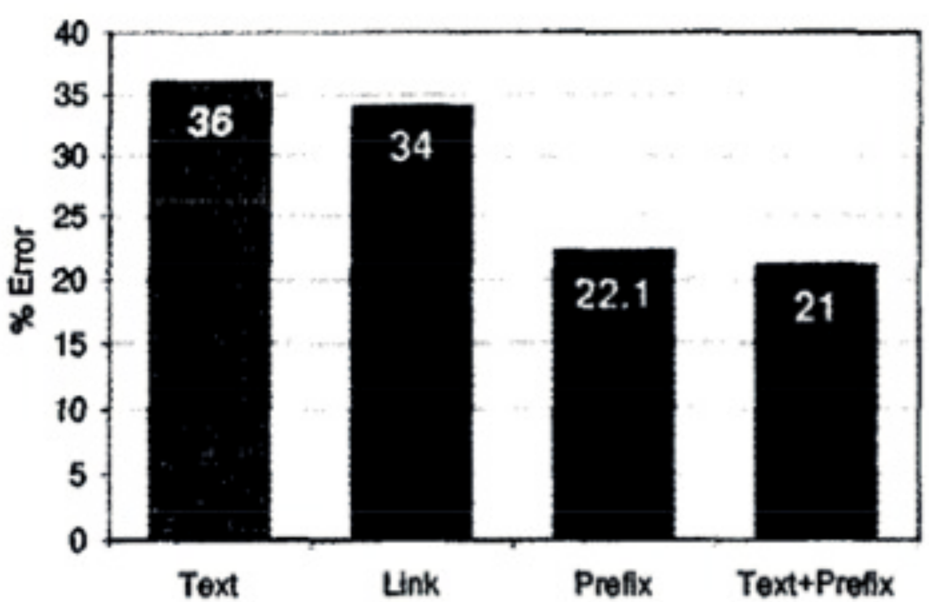
Object classification

Classification of web pages

[Chakrabarti, et al., SIGMOD98]

use pure text for classification: 36% error

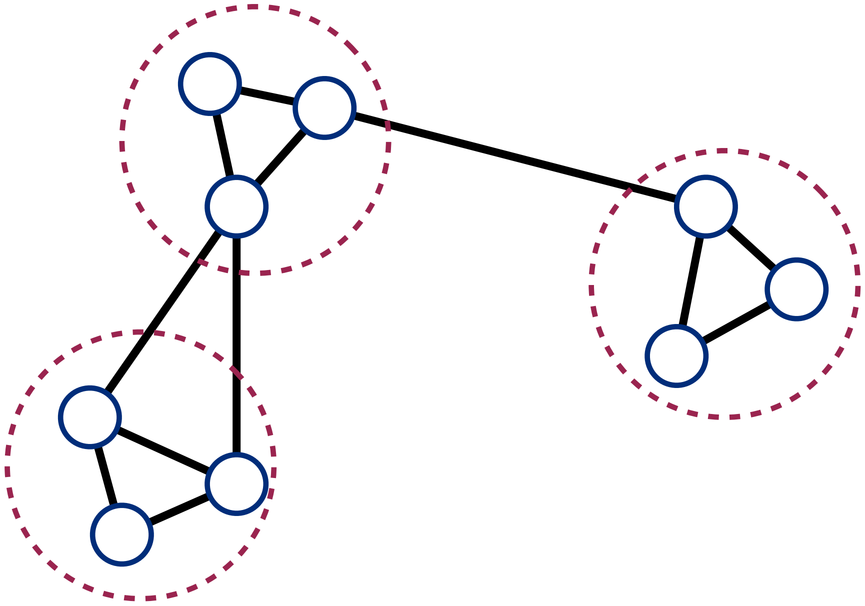
use neighbor predicted classes:
34% error, 22.1% error



Object clustering

Clustering nodes using link information

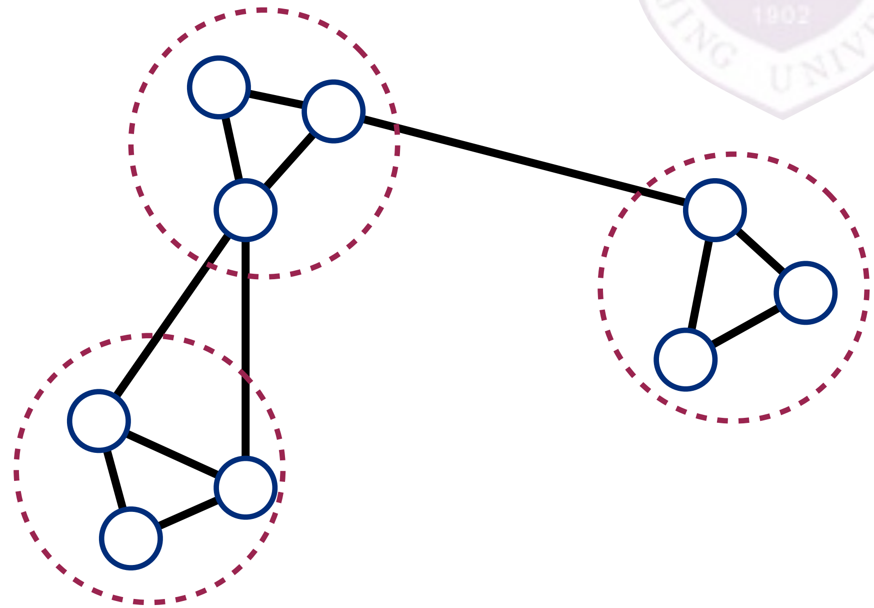
community discovery in social networks



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

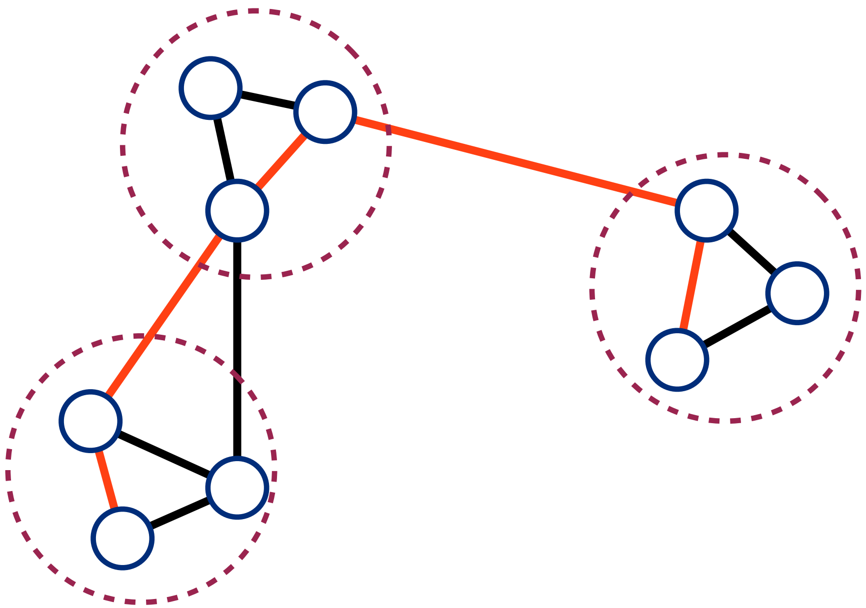
hierarchical clustering

graph-cut

k-medoids

Object clustering

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

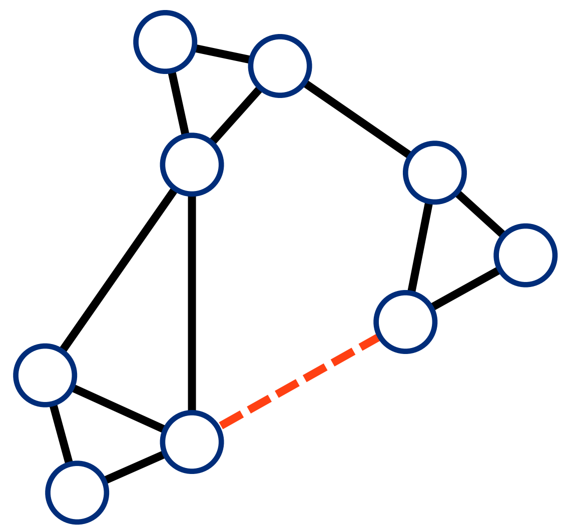


all clustering algorithms can be used

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]

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

$\Gamma(x)$ 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]

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

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

consider the relative counting

Link prediction

Similarities among nodes: neighbor-based

Preferential attachment [Mitzenmacher, ACCCC'01]

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

$\Gamma(x)$ 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]

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

$\mathsf{paths}_{x,y}^{\langle \ell \rangle}$ 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: $\mathsf{score}(x, y) = H_{x,y} + H_{y,x}$

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

normalized commute time:

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

π_x is the probability of x in the stationary distribution

Link prediction

Similarities among nodes: meta methods

SimRank [Jeh and Widom, KDD02]

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

recursively compute the similarity

习题

PageRank算法的思想是什么？