

KHAI THÁC DỮ LIỆU & KHAI PHÁ TRI THỨC

Data Mining & Knowledge Discovery

Bài 7. Gom cụm/ Clustering


Các phương pháp cơ bản

Mã MH: 505043

TS. HOÀNG Anh

Chapter 10. Cluster Analysis:

Khái niệm và Phương pháp cơ bản

- **Gom cụm: Khái niệm cơ bản** 
- Partitioning Methods
- Hierarchical Methods
- Density-Based Methods/ Các pp dựa trên mật độ
- Grid-Based Methods
- Đánh giá mô hình gom cụm
- Tổng kết

Cluster Analysis là gì?

- **Cụm/ Nhóm/ Cluster: Tập các đối tượng dữ liệu**
 - Tương đồng/ similar (related) trong cùng nhóm dữ liệu
 - Bất tương đồng/ dissimilar (unrelated) giữa các nhóm dữ liệu
- Gom cụm (cluster analysis, *clustering*, *data segmentation*, ...)
 - Tìm kiếm mức độ tương đồng giữa các điểm dữ liệu, tuân theo các đặc tính được tìm thấy và gom/ tập hợp các điểm dữ liệu tương đồng vào cùng nhóm/ clusters
- **Học không giám sát/ Unsupervised learning**: dữ liệu không nhãn (i.e., *learning by observations* vs. learning by examples: supervised)
- Ứng dụng cơ bản
 - **Stand-alone tool** -> Hiểu phân phối dữ liệu
 - **Preprocessing step** -> Các thuật toán còn lại

Gom cụm để hiểu dữ liệu và ứng dụng

- Biology: taxonomy of living things: kingdom, phylum, class, order, family, genus and species
- Information retrieval: document clustering
- Land use: Identification of areas of similar land use in an earth observation database
- Marketing: Help marketers discover distinct groups in their customer bases, and then use this knowledge to develop targeted marketing programs
- City-planning: Identifying groups of houses according to their house type, value, and geographical location
- Earth-quake studies: Observed earth quake epicenters should be clustered along continent faults
- Climate: understanding earth climate, find patterns of atmospheric and ocean
- Economic Science: market research

Gom cụm/ Công cụ xử lý (Utility)

- Tổng hợp/ Summarization:
 - Preprocessing for regression, PCA, classification, and association analysis
- Compression:
 - Image processing: vector quantization
- Thuật toán K-means
 - Localizing search to one or a small number of clusters
- Phát hiện điểm ngoại lai/ Outlier detection
 - Outliers are often viewed as those “far away” from any cluster

Chất lượng: Good Clustering?

- Một phương pháp gom cụm tốt sẽ tạo ra các nhóm/ cụm tốt
 - high intra-class similarity: **cohesive** within clusters
 - low inter-class similarity: **distinctive** between clusters
- Chất lượng của phương pháp gom cụm phụ thuộc vào
 - Phương pháp đo mức độ tương đồng
 - Thực nghiệm
 - Khả năng phát hiện mẫu

Đo lường chất lượng gom cụm

- **Chỉ số, độ đo Dissimilarity/Similarity**
 - Độ tương đồng được đo bởi hàm khoảng cách: $d(i, j)$
 - Định nghĩa hàm khoảng cách **distance functions** thường khác với interval-scaled, boolean, categorical, ordinal ratio, và vector variables
 - Các trọng số được tính từ nhiều biến tùy thuộc ứng dụng
- **Chất lượng gom cụm:**
 - There is usually a separate “quality” function that measures the “goodness” of a cluster.
 - It is hard to define “similar enough” or “good enough”
 - The answer is typically highly subjective

Gom cụm

- Chỉ số phân nhóm/ Partitioning criteria
 - Single level vs. hierarchical partitioning (often, multi-level hierarchical partitioning is desirable)
- Sự phân tách các nhóm/ Separation of clusters
 - Exclusive (e.g., one customer belongs to only one region) vs. non-exclusive (e.g., one document may belong to more than one class)
- Đo mức độ tương đồng/ Similarity measure
 - Distance-based (e.g., Euclidian, road network, vector) vs. connectivity-based (e.g., density or contiguity)
- Không gian cụm/ Clustering space
 - Full space (often when low dimensional) vs. subspaces (often in high-dimensional clustering)

Yêu cầu và Thách thức

Requirements and Challenges

- Scalability
 - Clustering all the data instead of only on samples
- Ability to deal with different types of attributes
 - Numerical, binary, categorical, ordinal, linked, and mixture of these
- Constraint-based clustering
 - User may give inputs on constraints
 - Use domain knowledge to determine input parameters
- Interpretability and usability
- Others
 - Discovery of clusters with arbitrary shape
 - Ability to deal with noisy data
 - Incremental clustering and insensitivity to input order
 - High dimensionality

Các phương pháp gom cụm cơ bản

Major Clustering Approaches (I)

- Partitioning approach:
 - Construct various partitions and then evaluate them by some criterion, e.g., minimizing the sum of square errors
 - Typical methods: **k-means**, k-medoids, CLARANS
- Hierarchical approach:
 - Create a hierarchical decomposition of the set of data (or objects) using some criterion
 - Typical methods: **Diana**, Agnes, BIRCH, CAMELEON
- Density-based approach:
 - Based on connectivity and density functions
 - Typical methods: **DBSACN**, OPTICS, DenClue
- Grid-based approach:
 - based on a multiple-level granularity structure
 - Typical methods: STING, WaveCluster, CLIQUE


Các phương pháp gom cụm cơ bản

Major Clustering Approaches (II)

- Model-based:
 - A model is hypothesized for each of the clusters and tries to find the best fit of that model to each other
 - Typical methods: EM, SOM, COBWEB
- Frequent pattern-based:
 - Based on the analysis of frequent patterns
 - Typical methods: p-Cluster
- User-guided or constraint-based:
 - Clustering by considering user-specified or application-specific constraints
 - Typical methods: COD (obstacles), constrained clustering
- Link-based clustering:
 - Objects are often linked together in various ways
 - Massive links can be used to cluster objects: SimRank, LinkClus

Chapter 10. Cluster Analysis:

Khái niệm và Phương pháp cơ bản

- Cluster Analysis: Basic Concepts
- **Các phương pháp Partitioning** 
- Hierarchical Methods
- Density-Based Methods
- Grid-Based Methods
- Evaluation of Clustering
- Summary

Thuật toán Partitioning: Khái niệm cơ bản

- Partitioning method: Partitioning a database D of n objects into a set of k clusters, such that the sum of squared distances is minimized (where c_i is the centroid or medoid of cluster C_i)

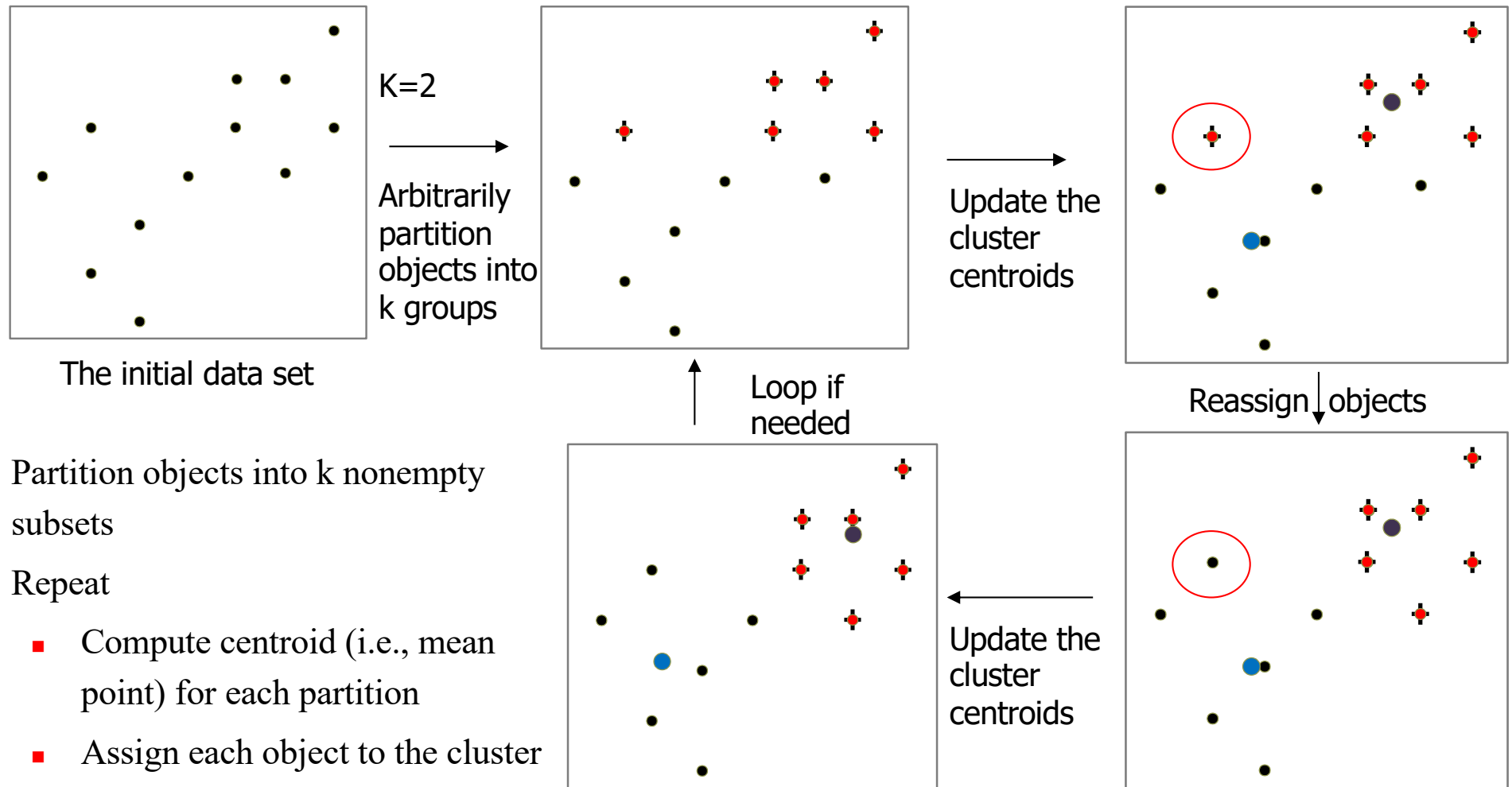
$$E = \sum_{i=1}^k \sum_{p \in C_i} (p - c_i)^2$$

- Given k , find a partition of k clusters that optimizes the chosen partitioning criterion
 - Global optimal: exhaustively enumerate all partitions
 - Heuristic methods: *k-means* and *k-medoids* algorithms
 - *k-means* (MacQueen'67, Lloyd'57/'82): Each cluster is represented by the center of the cluster
 - *k-medoids* or PAM (Partition around medoids) (Kaufman & Rousseeuw'87): Each cluster is represented by one of the objects in the cluster

Phương pháp *K-Means*

- Cho k , thuật toán *k-means* gồm 4 bước:
 - Phân nhóm dữ liệu thành k tập con không giao nhau
 - Tính điểm centroids của cụm/ nhóm hiện tại (the centroid is the center, i.e., *mean point*, of the cluster)
 - Gán mỗi điểm dữ liệu thuộc cụm có centroid gần nhất
 - Lặp lại bước 2, thuật toán dừng khi việc gán điểm dữ liệu “ổn định”

Ví dụ *K-Means* Clustering



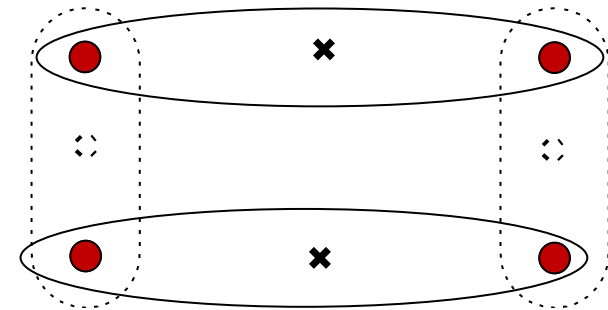
- Partition objects into k nonempty subsets
- Repeat
 - Compute centroid (i.e., mean point) for each partition
 - Assign each object to the cluster of its nearest centroid
- Until no change

Phương pháp *K-Means*

- Điểm mạnh: *Efficient*: $O(tkn)$, where n is # objects, k is # clusters, and t is # iterations. Normally, $k, t \ll n$.
 - So sánh: PAM: $O(k(n-k)^2)$, CLARA: $O(ks^2 + k(n-k))$
- Comment: thường đạt tối ưu cục bộ/ *local optimal*.
- Điểm yếu
 - Áp dụng được với không gian dữ liệu n -chiều, liên tục
 - Sử dụng k-modes với dữ liệu categorical
 - Sử dụng k-medoids với nhiều kiểu dữ liệu
 - Cần xác định k , số lượng cụm, (there are ways to automatically determine the best k (see Hastie et al., 2009))
 - Nhạy cảm với dữ liệu nhiễu và *outliers*
 - Not suitable to discover clusters with *non-convex shapes*

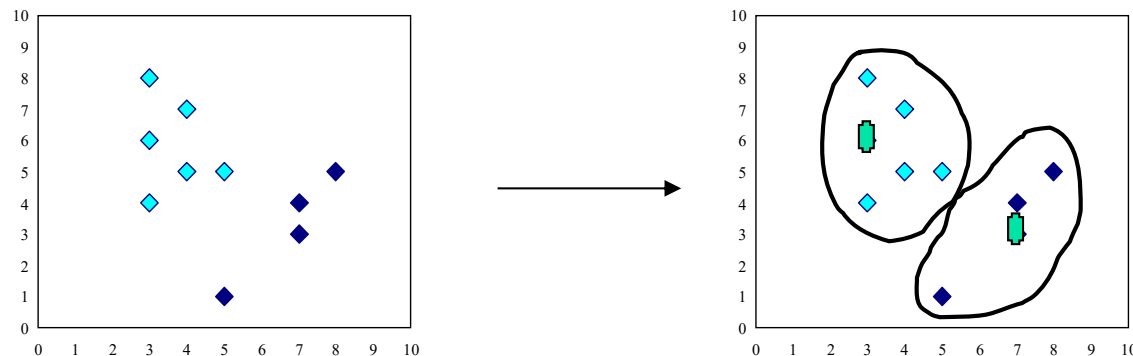
Biến thể K -Means

- Các biến thể k -means khác nhau ở
 - Lựa chọn giá trị k ban đầu
 - Tính toán độ tương đồng/ bất tương đồng
 - Phương pháp tính cluster means
- Xử lý dữ liệu categorical: k -modes
 - Thay thế trị trung bình của cụm với modes
 - Sử dụng các phép đo độ bất tương đồng với dữ liệu categorical
 - Using a frequency-based method to update modes of clusters
 - A mixture of categorical and numerical data: k -prototype method

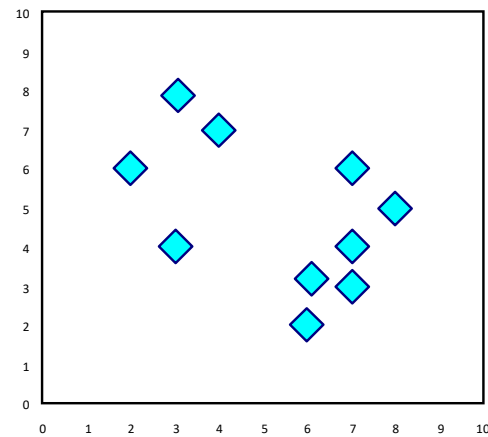


Vấn đề với K-Means?

- Thuật toán K-Means nhạy cảm với outliers!
 - Since an object with an extremely large value may substantially distort the distribution of the data
- K-Medoids: Instead of taking the **mean** value of the object in a cluster as a reference point, **medoids** can be used, which is the **most centrally located** object in a cluster

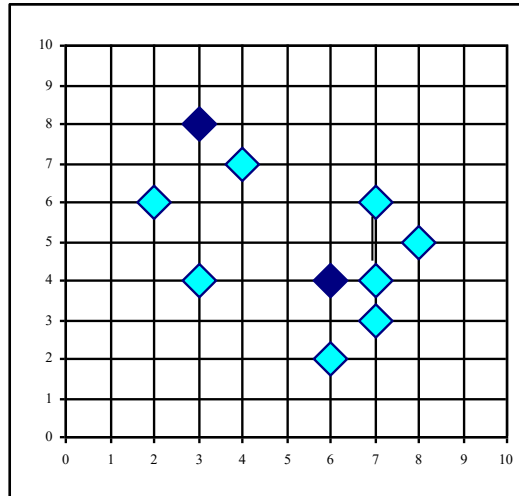


PAM: A Typical K-Medoids Algorithm

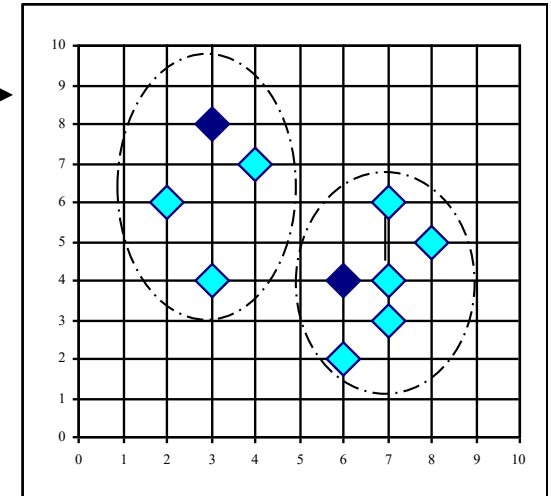


K=2

Arbitrary
choose k
object as
initial
medoids



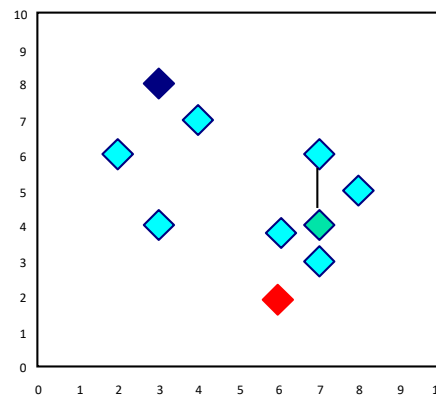
Assign
each
remainin
g object
to
nearest
medoids



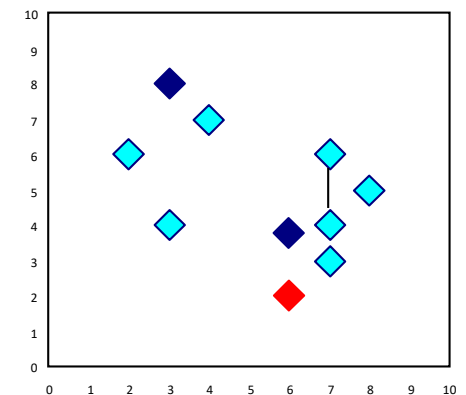
Randomly select a
nonmedoid object, O_{random}

**Do loop
Until no
change**

Swapping O
and O_{random}
If quality is
improved.



Compute
total cost of
swapping




Phương pháp K-Medoid

- *K-Medoids* Clustering: Find *representative* objects (medoids) in clusters
 - *PAM* (Partitioning Around Medoids, Kaufmann & Rousseeuw 1987)
 - Starts from an initial set of medoids and iteratively replaces one of the medoids by one of the non-medoids if it improves the total distance of the resulting clustering
 - *PAM* works effectively for small data sets, but does not scale well for large data sets (due to the computational complexity)
- Efficiency improvement on PAM
 - *CLARA* (Kaufmann & Rousseeuw, 1990): PAM on samples
 - *CLARANS* (Ng & Han, 1994): Randomized re-sampling

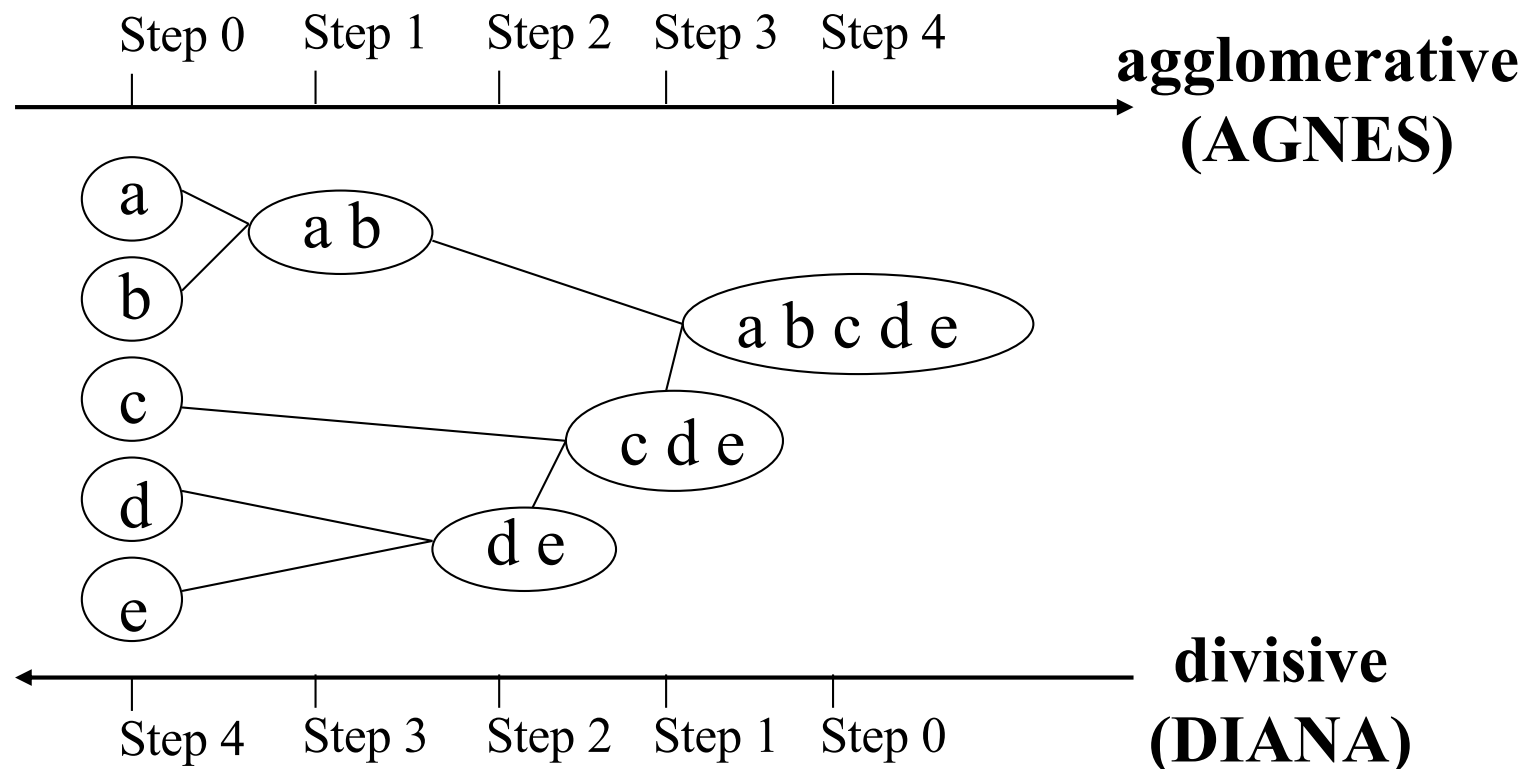
Chapter 10. Cluster Analysis:

Khái niệm và Phương pháp cơ bản

- Cluster Analysis: Basic Concepts
- Partitioning Methods
- **Phương pháp Hierarchical** 
- Density-Based Methods
- Grid-Based Methods
- Evaluation of Clustering
- Summary

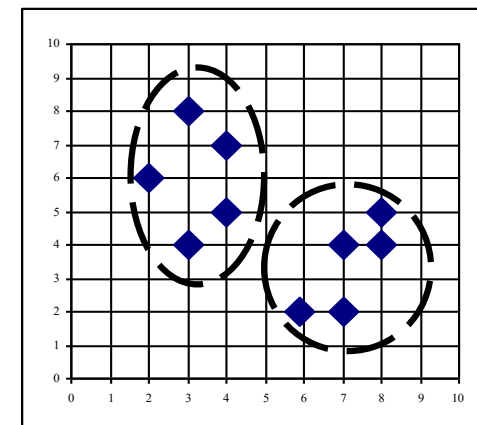
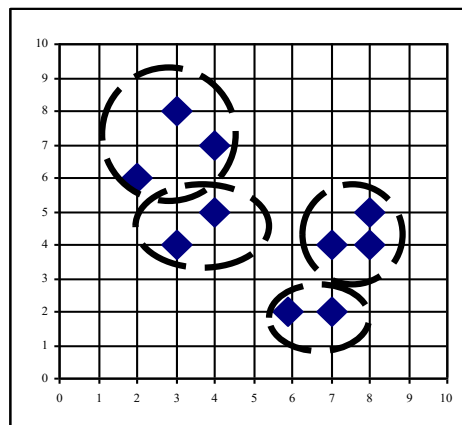
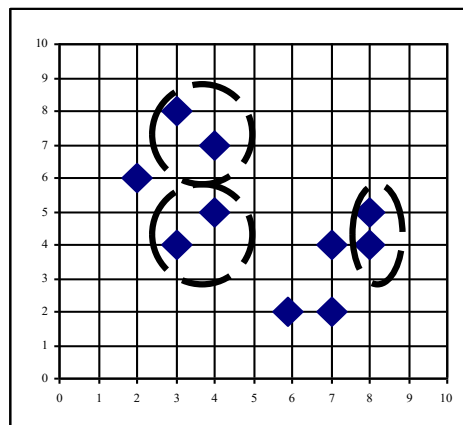
Hierarchical Clustering

- Sử dụng ma trận khoảng cách. Phương pháp này không yêu cầu số lượng cụm k là đầu vào, nhưng cần điều kiện kết thúc.

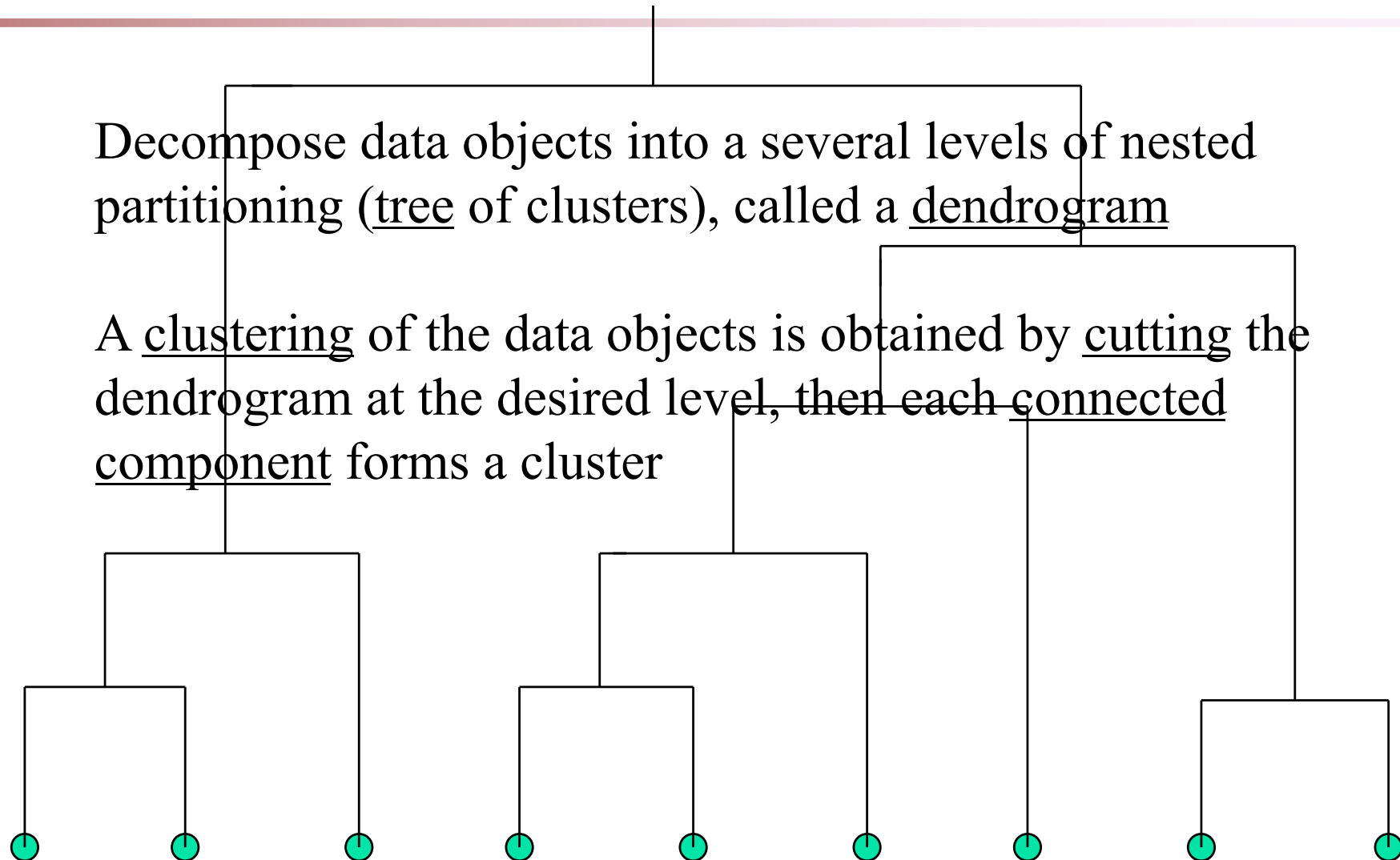


AGNES (Agglomerative Nesting)

- Introduced in Kaufmann and Rousseeuw (1990)
- Implemented in statistical packages, e.g., Splus
- Use the **single-link** method and the dissimilarity matrix
- Merge nodes that have the least dissimilarity
- Go on in a non-descending fashion
- Eventually all nodes belong to the same cluster

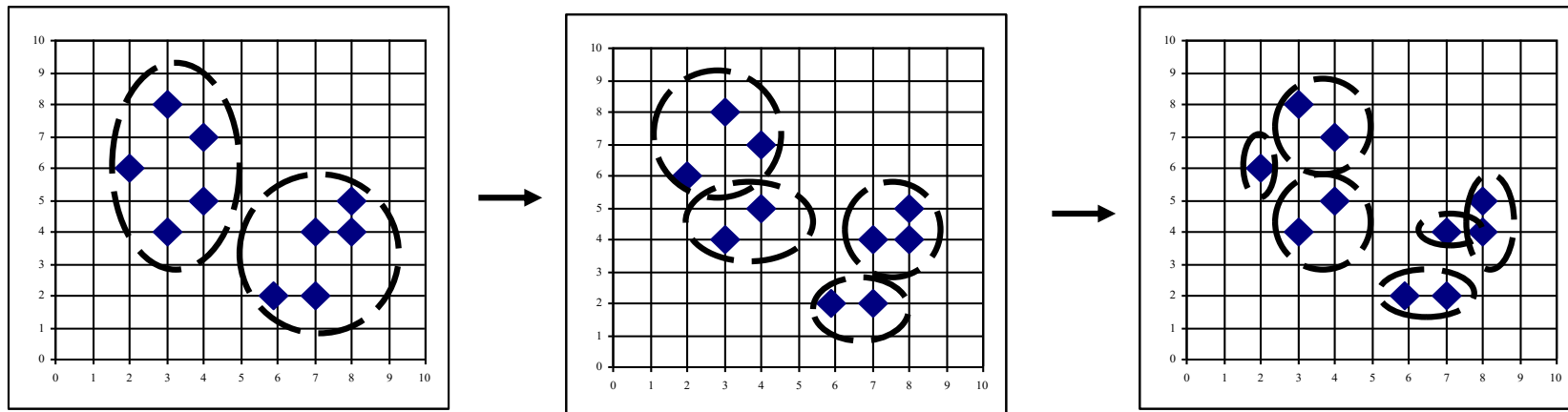


Dendrogram: Shows How Clusters are Merged

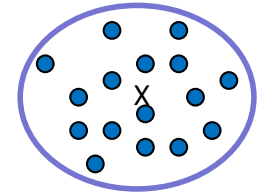
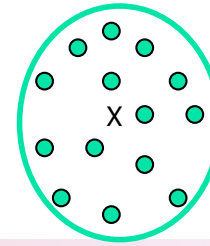


DIANA (Divisive Analysis)

- Introduced in Kaufmann and Rousseeuw (1990)
- Implemented in statistical analysis packages, e.g., Splus
- Inverse order of AGNES
- Eventually each node forms a cluster on its own



Distance between Clusters



- Single link: smallest distance between an element in one cluster and an element in the other, i.e., $\text{dist}(K_i, K_j) = \min(t_{ip}, t_{jq})$
- Complete link: largest distance between an element in one cluster and an element in the other, i.e., $\text{dist}(K_i, K_j) = \max(t_{ip}, t_{jq})$
- Average: avg distance between an element in one cluster and an element in the other, i.e., $\text{dist}(K_i, K_j) = \text{avg}(t_{ip}, t_{jq})$
- Centroid: distance between the centroids of two clusters, i.e., $\text{dist}(K_i, K_j) = \text{dist}(C_i, C_j)$
- Medoid: distance between the medoids of two clusters, i.e., $\text{dist}(K_i, K_j) = \text{dist}(M_i, M_j)$
 - Medoid: a chosen, centrally located object in the cluster

Centroid, Radius and Diameter of a Cluster (for numerical data sets)

- Centroid: the “middle” of a cluster

$$C_m = \frac{\sum_{i=1}^N (t_{ip})}{N}$$

- Radius: square root of average distance from any point of the cluster to its centroid

$$R_m = \sqrt{\frac{\sum_{i=1}^N (t_{ip} - c_m)^2}{N}}$$

- Diameter: square root of average mean squared distance between all pairs of points in the cluster

$$D_m = \sqrt{\frac{\sum_{i=1}^N \sum_{i=1}^N (t_{ip} - t_{iq})^2}{N(N-1)}}$$

Extensions to Hierarchical Clustering

- Major weakness of agglomerative clustering methods
 - Can never undo what was done previously
 - Do not scale well: time complexity of at least $O(n^2)$, where n is the number of total objects
- Integration of hierarchical & distance-based clustering
 - BIRCH (1996): uses CF-tree and incrementally adjusts the quality of sub-clusters
 - CHAMELEON (1999): hierarchical clustering using dynamic modeling

BIRCH (Balanced Iterative Reducing and Clustering Using Hierarchies)

- Zhang, Ramakrishnan & Livny, SIGMOD'96
- Incrementally construct a CF (Clustering Feature) tree, a hierarchical data structure for multiphase clustering
 - Phase 1: scan DB to build an initial in-memory CF tree (a multi-level compression of the data that tries to preserve the inherent clustering structure of the data)
 - Phase 2: use an arbitrary clustering algorithm to cluster the leaf nodes of the CF-tree
- *Scales linearly*: finds a good clustering with a single scan and improves the quality with a few additional scans
- *Weakness*: handles only numeric data, and sensitive to the order of the data record

Clustering Feature Vector in BIRCH

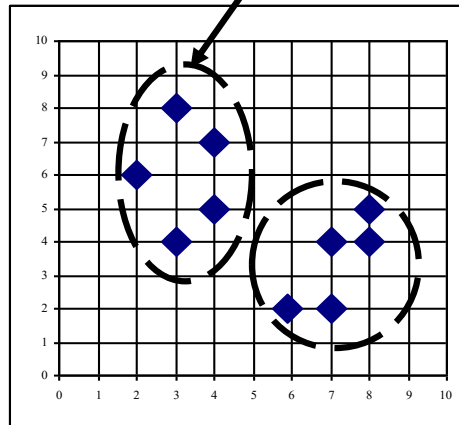
Clustering Feature (CF): $CF = (N, LS, SS)$

N : Number of data points

LS : linear sum of N points: $\sum_{i=1}^N X_i$

SS : square sum of N points

$$\sum_{i=1}^N X_i^2$$



$CF = (5, (16,30), (54,190))$

(3,4)

(2,6)

(4,5)

(4,7)

(3,8)

CF-Tree in BIRCH

- Clustering feature:
 - Summary of the statistics for a given subcluster: the 0-th, 1st, and 2nd moments of the subcluster from the statistical point of view
 - Registers crucial measurements for computing cluster and utilizes storage efficiently
- A CF tree is a height-balanced tree that stores the clustering features for a hierarchical clustering
 - A nonleaf node in a tree has descendants or “children”
 - The nonleaf nodes store sums of the CFs of their children
- A CF tree has two parameters
 - Branching factor: max # of children
 - Threshold: max diameter of sub-clusters stored at the leaf nodes

The CF Tree Structure

Root

$B = 7$

$L = 6$

| | | | | |
|--------------------|--------------------|--------------------|-------|--------------------|
| CF ₁ | CF ₂ | CF ₃ | | CF ₆ |
| child ₁ | child ₂ | child ₃ | | child ₆ |

Non-leaf node

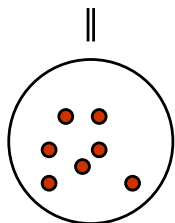
| | | | | |
|--------------------|--------------------|--------------------|-------|--------------------|
| CF ₁ | CF ₂ | CF ₃ | | CF ₅ |
| child ₁ | child ₂ | child ₃ | | child ₅ |

Leaf node

Leaf node

| | | | | | |
|------|-----------------|-----------------|-------|-----------------|------|
| prev | CF ₁ | CF ₂ | | CF ₆ | next |
|------|-----------------|-----------------|-------|-----------------|------|

| | | | | | |
|------|-----------------|-----------------|-------|-----------------|------|
| prev | CF ₁ | CF ₂ | | CF ₄ | next |
|------|-----------------|-----------------|-------|-----------------|------|



The Birch Algorithm

- Cluster Diameter

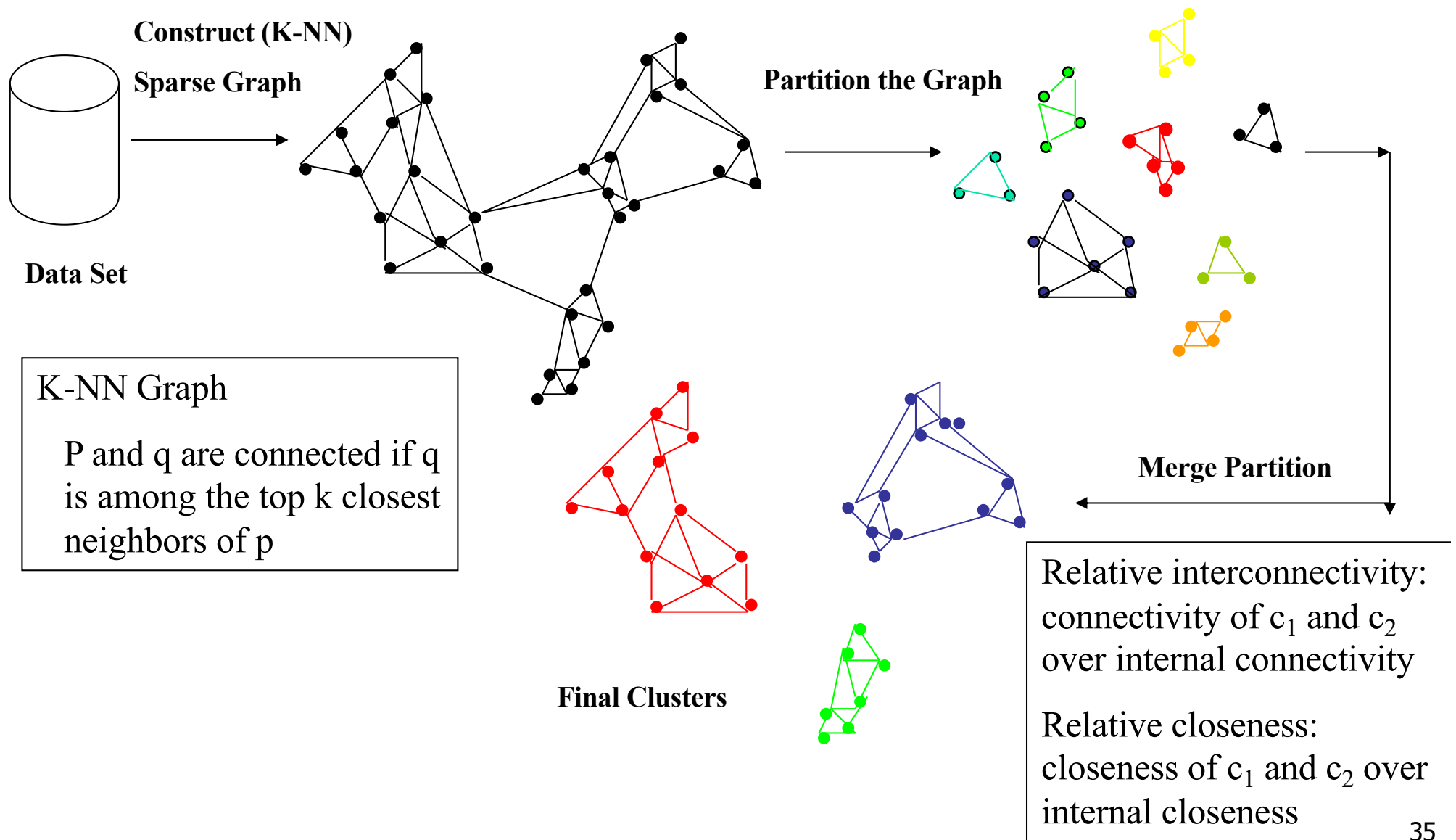
$$\sqrt{\frac{1}{n(n-1)} \sum (x_i - x_j)^2}$$

- For each point in the input
 - Find closest leaf entry
 - Add point to leaf entry and update CF
 - If entry diameter $>$ max_diameter, then split leaf, and possibly parents
- Algorithm is $O(n)$
- Concerns
 - Sensitive to insertion order of data points
 - Since we fix the size of leaf nodes, so clusters may not be so natural
 - Clusters tend to be spherical given the radius and diameter measures

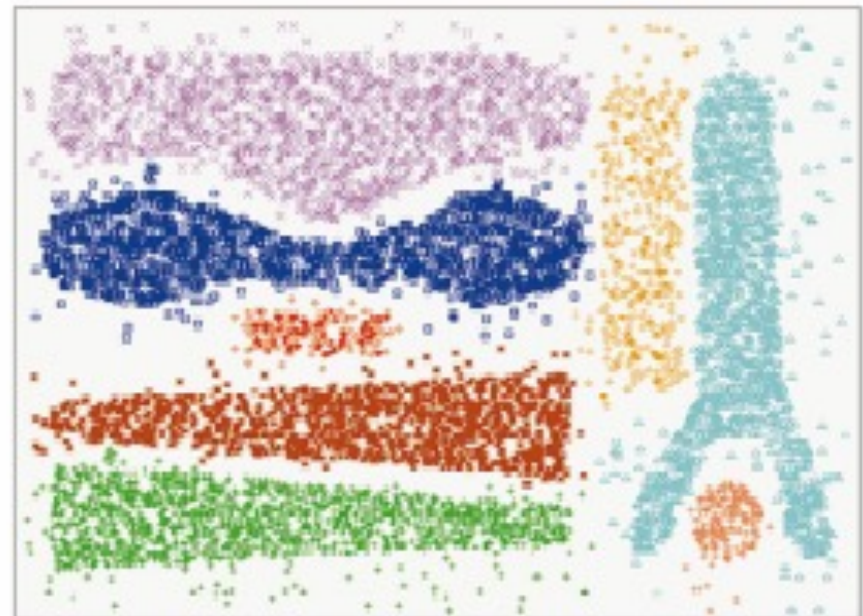
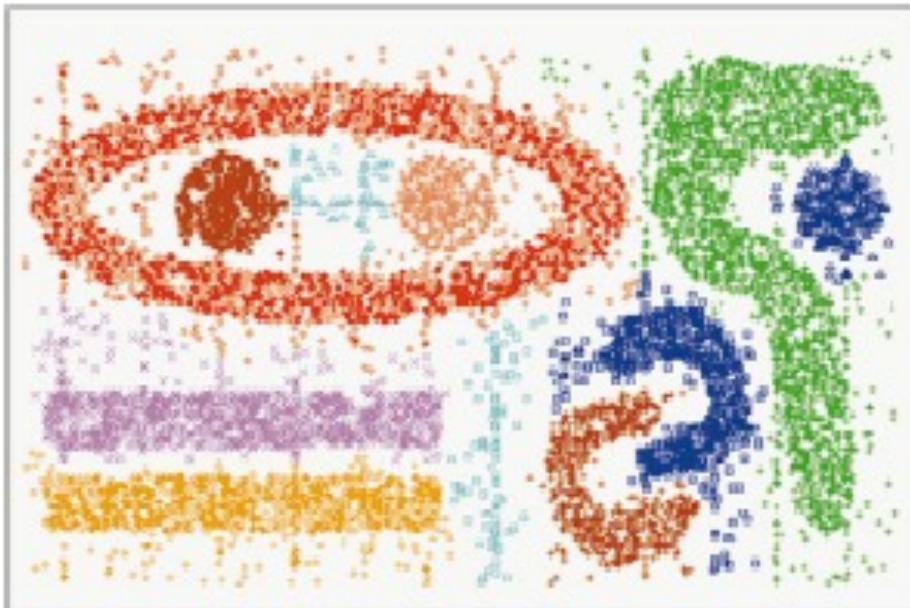
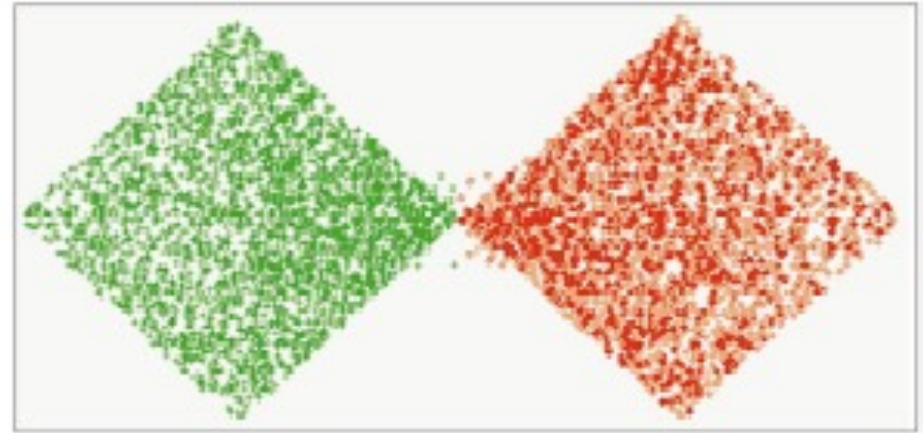
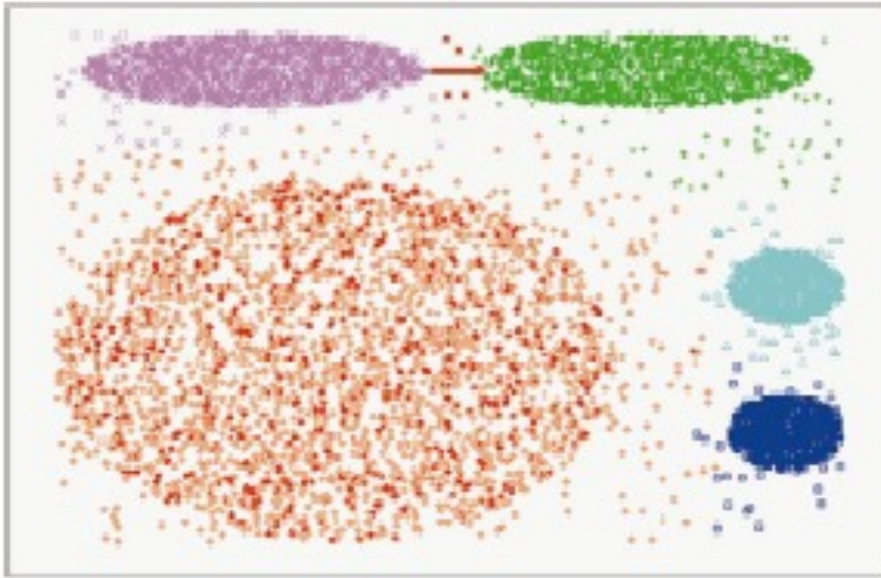
CHAMELEON: Hierarchical Clustering Using Dynamic Modeling (1999)

- CHAMELEON: G. Karypis, E. H. Han, and V. Kumar, 1999
- Measures the similarity based on a dynamic model
 - Two clusters are merged only if the *interconnectivity* and *closeness (proximity)* between two clusters are high *relative to* the internal interconnectivity of the clusters and closeness of items within the clusters
- Graph-based, and a two-phase algorithm
 1. Use a graph-partitioning algorithm: cluster objects into a large number of relatively small sub-clusters
 2. Use an agglomerative hierarchical clustering algorithm: find the genuine clusters by repeatedly combining these sub-clusters

Overall Framework of CHAMELEON



CHAMELEON (Clustering Complex Objects)



Probabilistic Hierarchical Clustering

- Algorithmic hierarchical clustering
 - Nontrivial to choose a good distance measure
 - Hard to handle missing attribute values
 - Optimization goal not clear: heuristic, local search
- Probabilistic hierarchical clustering
 - Use probabilistic models to measure distances between clusters
 - Generative model: Regard the set of data objects to be clustered as a sample of the underlying data generation mechanism to be analyzed
 - Easy to understand, same efficiency as algorithmic agglomerative clustering method, can handle partially observed data
- In practice, assume the generative models adopt common distributions functions, e.g., Gaussian distribution or Bernoulli distribution, governed by parameters

Generative Model

- Given a set of 1-D points $X = \{x_1, \dots, x_n\}$ for clustering analysis & assuming they are generated by a Gaussian distribution:

$$\mathcal{N}(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- The probability that a point $x_i \in X$ is generated by the model

$$P(x_i | \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i - \mu)^2}{2\sigma^2}}$$

- The likelihood that X is generated by the model:

$$L(\mathcal{N}(\mu, \sigma^2) : X) = P(X | \mu, \sigma^2) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i - \mu)^2}{2\sigma^2}}$$

- The task of learning the generative model: find the parameters μ and σ^2 such that

$$\mathcal{N}(\mu_0, \sigma_0^2) = \arg \max \{ L(\mathcal{N}(\mu, \sigma^2) : X) \}$$

the maximum likelihood

A Probabilistic Hierarchical Clustering Algorithm

- For a set of objects partitioned into m clusters C_1, \dots, C_m , the quality can be measured by,

$$Q(\{C_1, \dots, C_m\}) = \prod_{i=1}^m P(C_i)$$

where $P()$ is the maximum likelihood

- Distance between clusters C_1 and C_2 : $dist(C_i, C_j) = -\log \frac{P(C_1 \cup C_2)}{P(C_1)P(C_2)}$
- Algorithm: Progressively merge points and clusters

Input: $D = \{o_1, \dots, o_n\}$: a data set containing n objects

Output: A hierarchy of clusters

Method

Create a cluster for each object $C_i = \{o_i\}$, $1 \leq i \leq n$;

For $i = 1$ to n {


Find pair of clusters C_i and C_j such that

$$C_i, C_j = \operatorname{argmax}_{i \neq j} \{\log (P(C_i \cup C_j) / (P(C_i)P(C_j)))\};$$

If $\log (P(C_i \cup C_j) / (P(C_i)P(C_j))) > 0$ then merge C_i and C_j }

Chapter 10. Cluster Analysis:

Khái niệm và Phương pháp cơ bản

- Cluster Analysis: Basic Concepts
- Partitioning Methods
- Hierarchical Methods
- **Phương pháp Density-Based** 
- Grid-Based Methods
- Evaluation of Clustering
- Summary

Phương pháp gom cụm dựa theo mật độ

Density-Based Clustering Methods

- Gom cụm theo mật độ (local cluster criterion), như là các điểm được kết nối theo mật độ
- Đặc tính cơ bản:
 - Discover clusters of arbitrary shape
 - Handle noise
 - One scan
 - Need density parameters as termination condition
- Thuật toán:
 - DBSCAN: Ester, et al. (KDD'96)
 - OPTICS: Ankerst, et al (SIGMOD'99).
 - DENCLUE: Hinneburg & D. Keim (KDD'98)
 - CLIQUE: Agrawal, et al. (SIGMOD'98) (more grid-based)

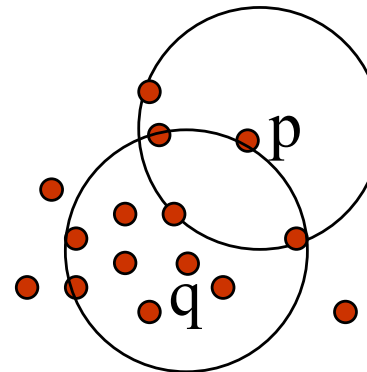
Gom cụm theo mật độ: Các khái niệm cơ bản

Density-Based Clustering: Basic Concepts

- Tham số cơ bản:
 - *Eps*: Maximum radius of the neighbourhood
 - *MinPts*: Minimum number of points in an Eps-neighbourhood of that point
- $N_{Eps}(p)$: $\{q \text{ belongs to } D \mid \text{dist}(p,q) \leq Eps\}$
- **Directly density-reachable**: A point p is directly density-reachable from a point q w.r.t. Eps , $MinPts$ if

- p belongs to $N_{Eps}(q)$
- core point condition:

$$|N_{Eps}(q)| \geq MinPts$$



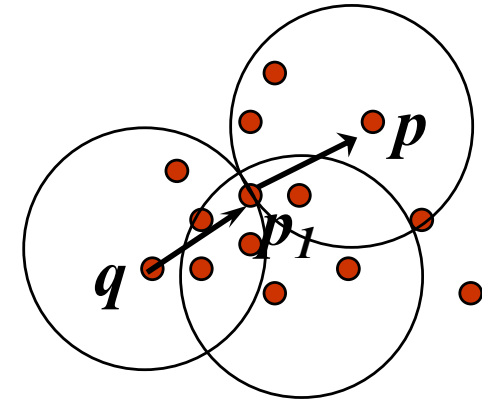
$MinPts = 5$

$Eps = 1 \text{ cm}$

Density-Reachable and Density-Connected

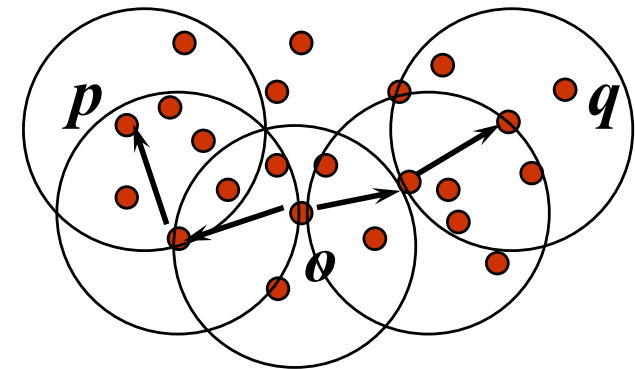
- Density-reachable:

- A point p is **density-reachable** from a point q w.r.t. Eps , $MinPts$ if there is a chain of points $p_1, \dots, p_n, p_1 = q, p_n = p$ such that p_{i+1} is directly density-reachable from p_i



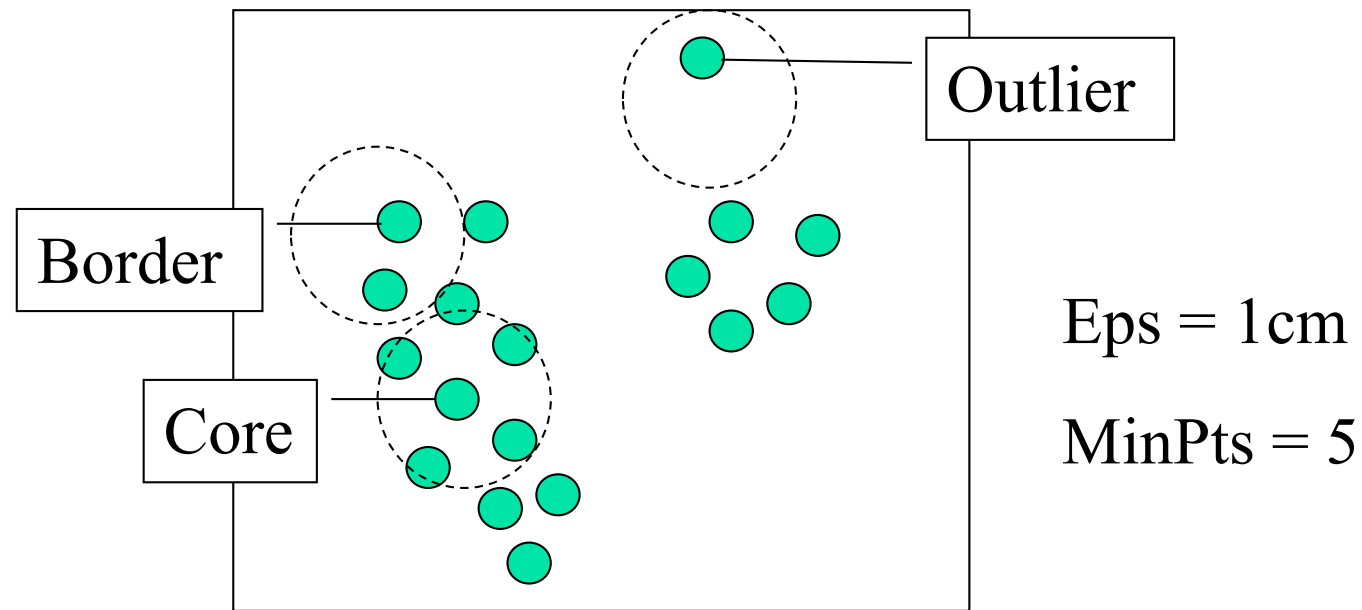
- Density-connected

- A point p is **density-connected** to a point q w.r.t. Eps , $MinPts$ if there is a point o such that both, p and q are density-reachable from o w.r.t. Eps and $MinPts$



DBSCAN: Density-Based Spatial Clustering of Applications with Noise

- Relies on a *density-based* notion of cluster: A *cluster* is defined as a maximal set of density-connected points
- Discovers clusters of arbitrary shape in spatial databases with noise



DBSCAN: Giải thuật

- Arbitrary select a point p
- Retrieve all points density-reachable from p w.r.t. Eps and $MinPts$
- If p is a core point, a cluster is formed
- If p is a border point, no points are density-reachable from p and DBSCAN visits the next point of the database
- Continue the process until all of the points have been processed

DBSCAN: Nhạy với các tham số

Figure 8. DBScan results for DS1 with MinPts at 4 and Eps at (a) 0.5 and (b) 0.4.

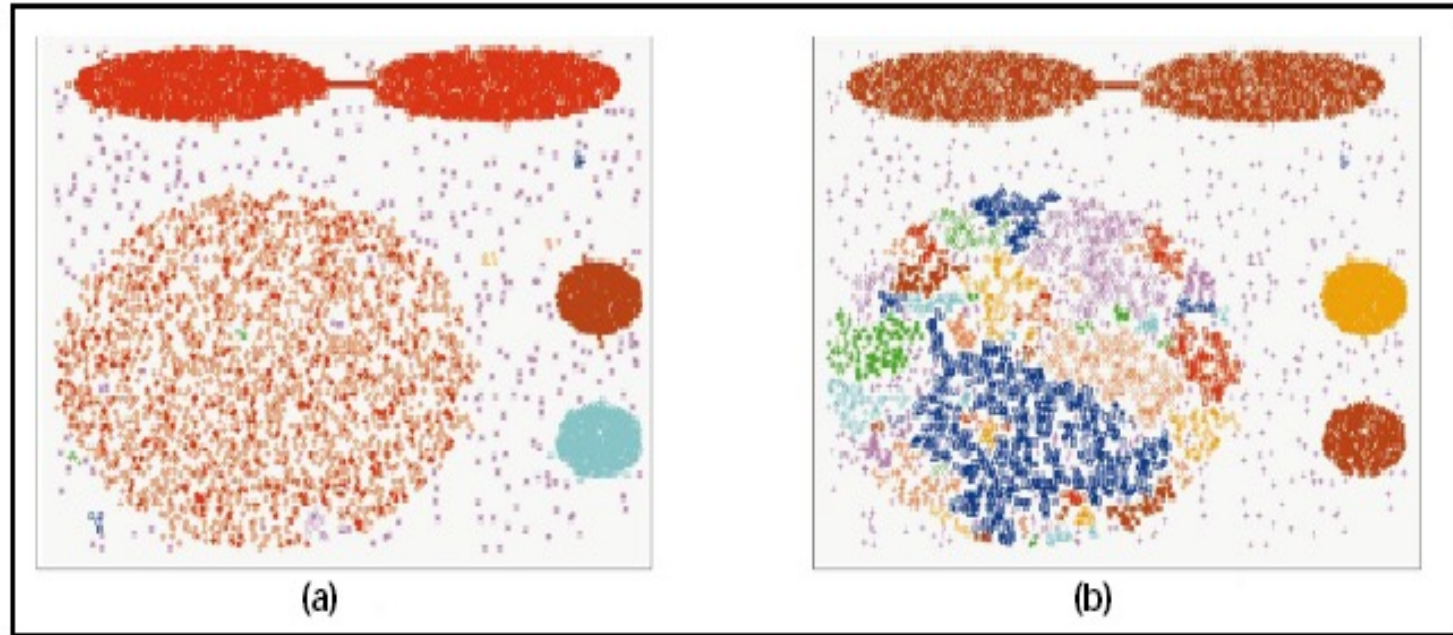
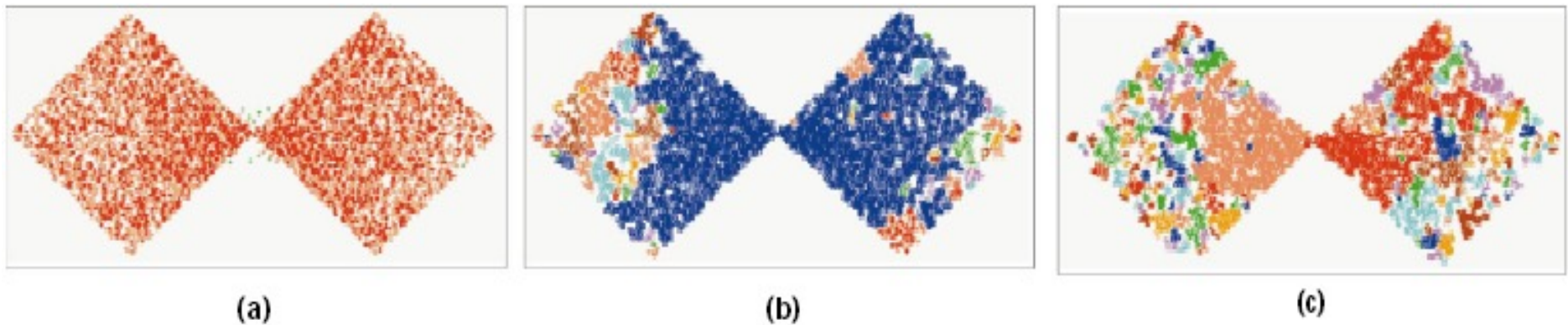


Figure 9. DBScan results for DS2 with MinPts at 4 and Eps at (a) 5.0, (b) 3.5, and (c) 3.0.



OPTICS: A Cluster-Ordering Method (1999)

- OPTICS: Ordering Points To Identify the Clustering Structure
 - Ankerst, Breunig, Kriegel, and Sander (SIGMOD'99)
 - Produces a special order of the database wrt its density-based clustering structure
 - This cluster-ordering contains info equiv to the density-based clusterings corresponding to a broad range of parameter settings
 - Good for both automatic and interactive cluster analysis, including finding intrinsic clustering structure
 - Can be represented graphically or using visualization techniques

OPTICS: Mở rộng từ DBSCAN

- Index-based:

- k = number of dimensions

- $N = 20$

- $p = 75\%$

- $M = N(1-p) = 5$

- Complexity: $O(N \log N)$

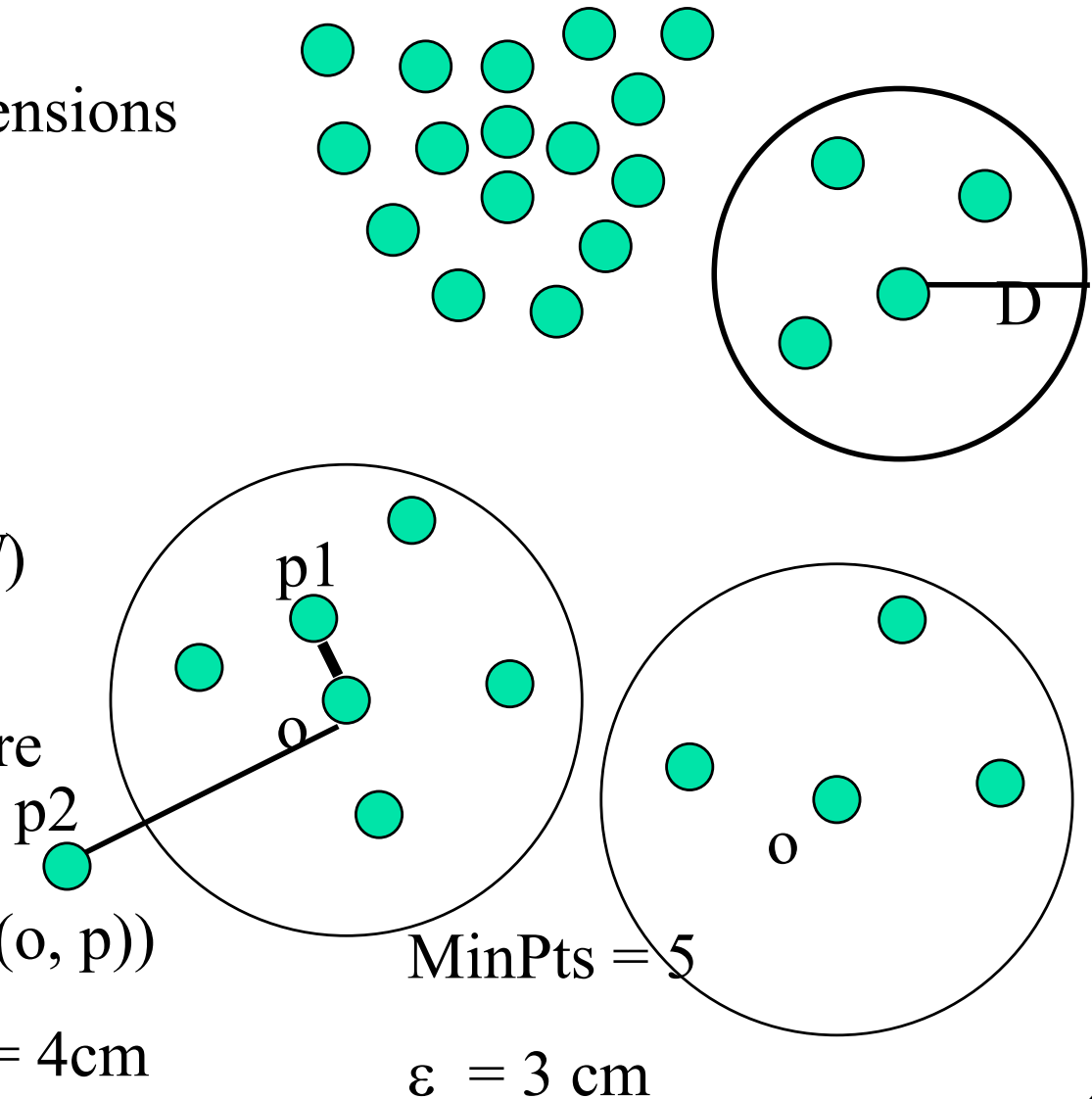
- Core Distance:

- min eps s.t. point is core

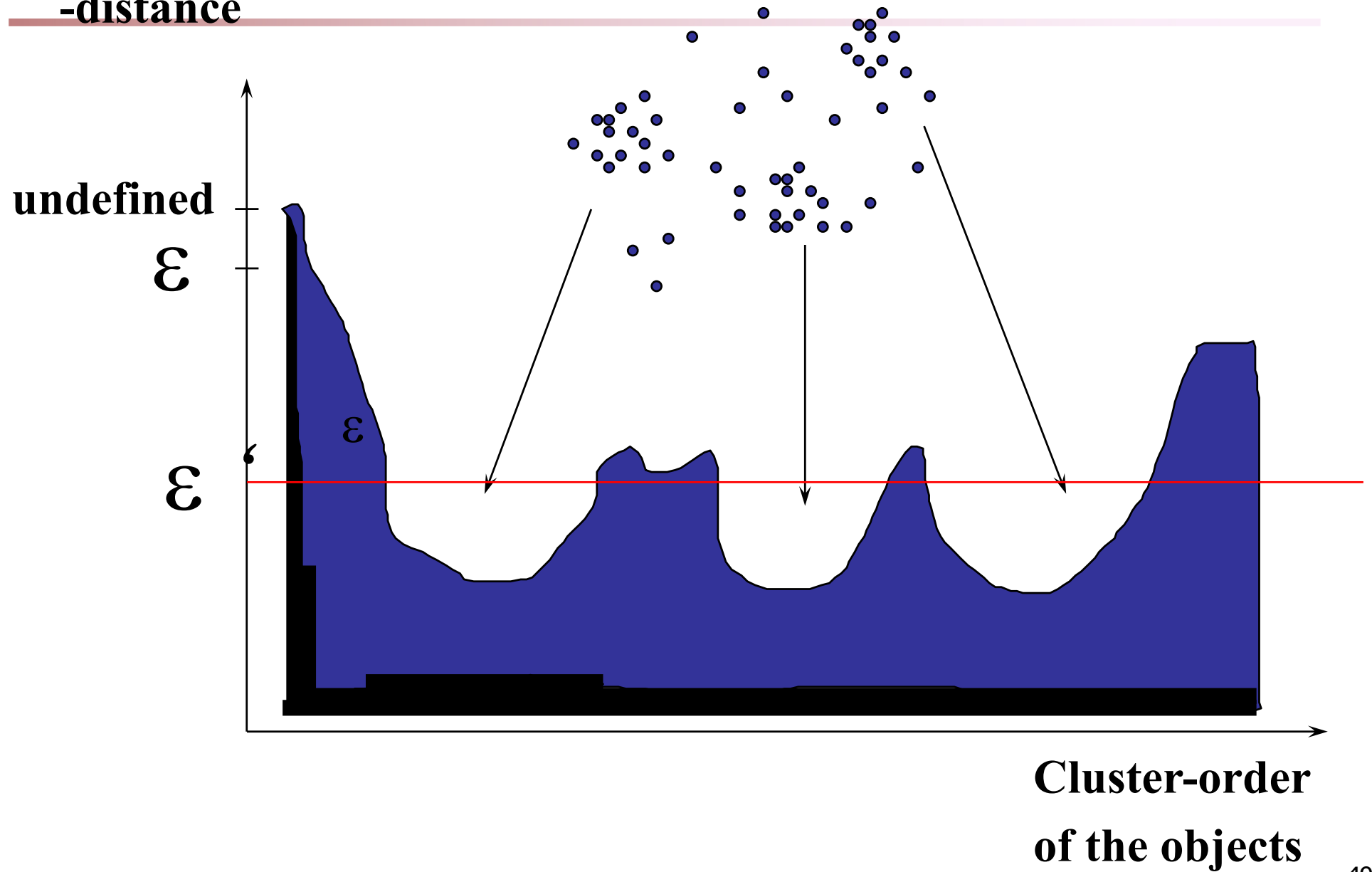
- Reachability Distance

$\text{Max}(\text{core-distance}(o), d(o, p))$

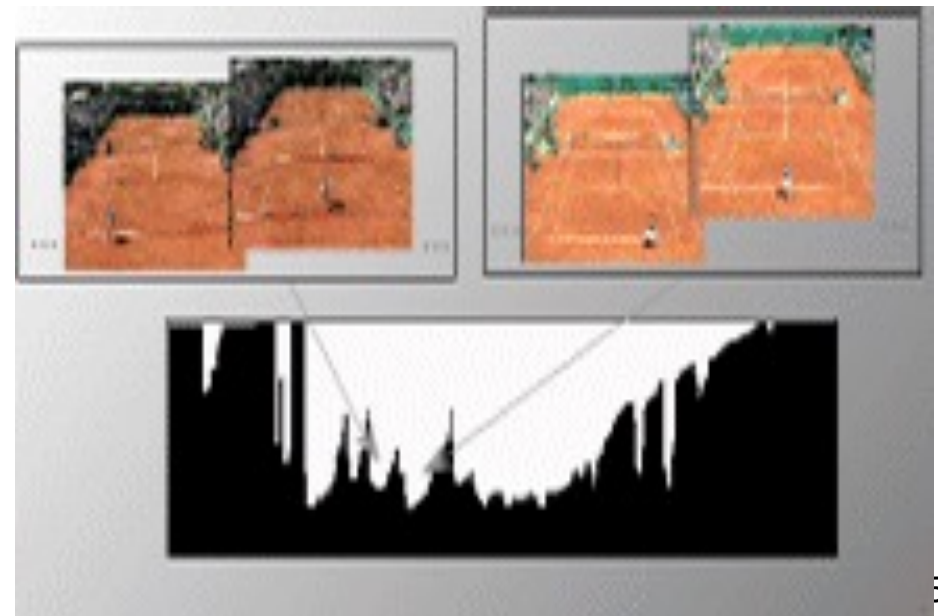
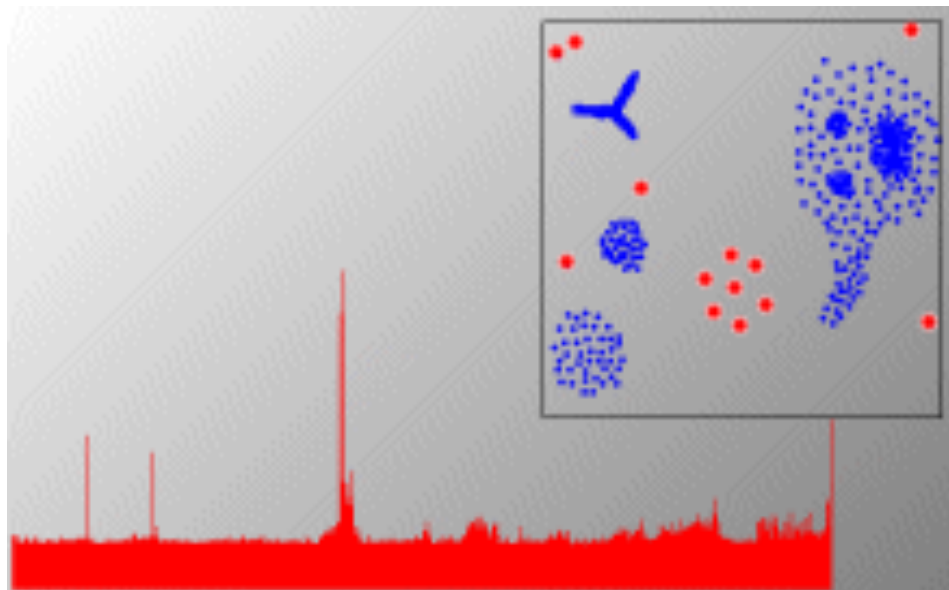
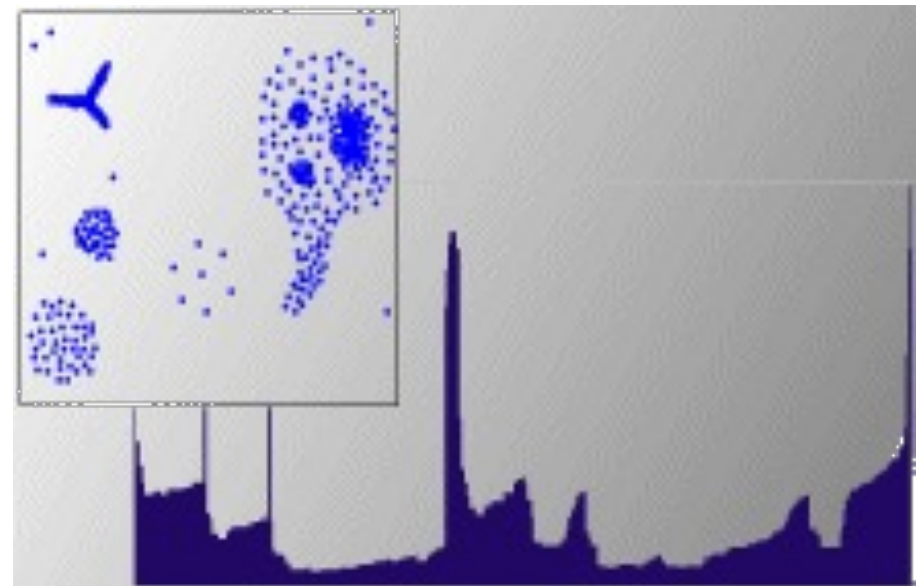
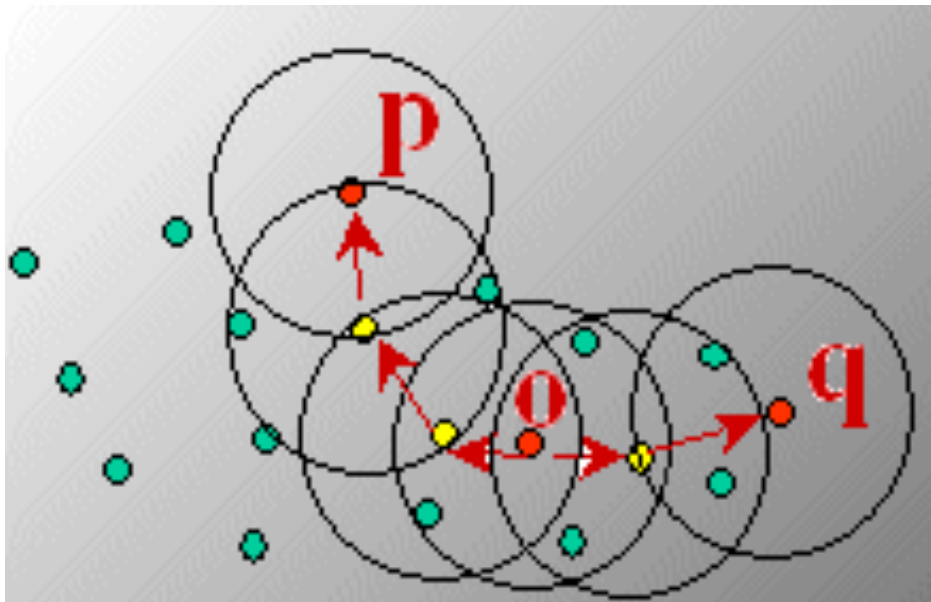
$r(p1, o) = 2.8\text{cm}$. $r(p2, o) = 4\text{cm}$



Reachability -distance



Density-Based Clustering: OPTICS & Its Applications



DENCLUE: Using Statistical Density Functions

- DENSity-based CLUstEring by Hinneburg & Keim (KDD'98)
- Using statistical density functions:

$$f_{Gaussian}(x, y) = e^{-\frac{d(x, y)^2}{2\sigma^2}}$$

influence of y
on x

$$f_{Gaussian}^D(x) = \sum_{i=1}^N e^{-\frac{d(x, x_i)^2}{2\sigma^2}}$$

total influence
on x

- Major features

$$\nabla f_{Gaussian}^D(x, x_i) = \sum_{i=1}^N (x_i - x) \cdot e^{-\frac{d(x, x_i)^2}{2\sigma^2}}$$

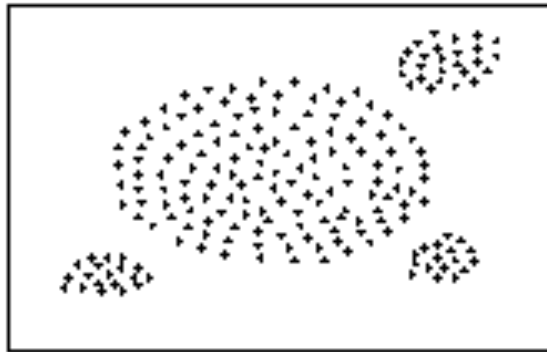
gradient of x in
the direction of
 x_i

- Solid mathematical foundation
- Good for data sets with large amounts of noise
- Allows a compact mathematical description of arbitrarily shaped clusters in high-dimensional data sets
- Significant faster than existing algorithm (e.g., DBSCAN)
- But needs a large number of parameters

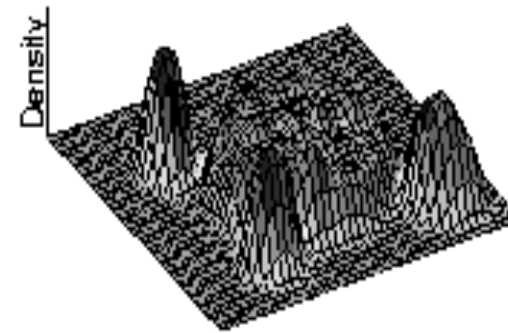
Denclue: Technical Essence

- Uses grid cells but only keeps information about grid cells that do actually contain data points and manages these cells in a tree-based access structure
- Influence function: describes the impact of a data point within its neighborhood
- Overall density of the data space can be calculated as the sum of the influence function of all data points
- Clusters can be determined mathematically by identifying density attractors
- Density attractors are local maximal of the overall density function
- Center defined clusters: assign to each density attractor the points density attracted to it
- Arbitrary shaped cluster: merge density attractors that are connected through paths of high density ($>$ threshold)

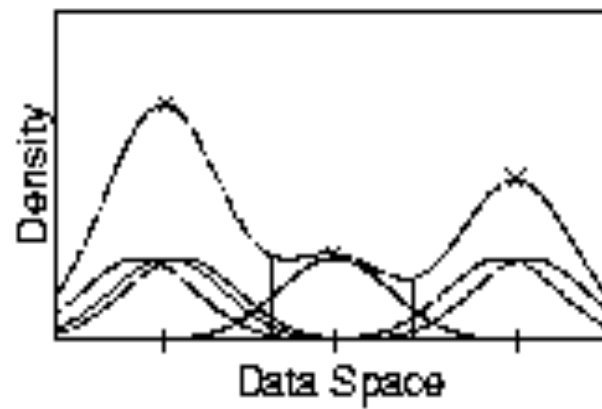
Density Attractor



(a) Data Set



(c) Gaussian



Center-Defined and Arbitrary

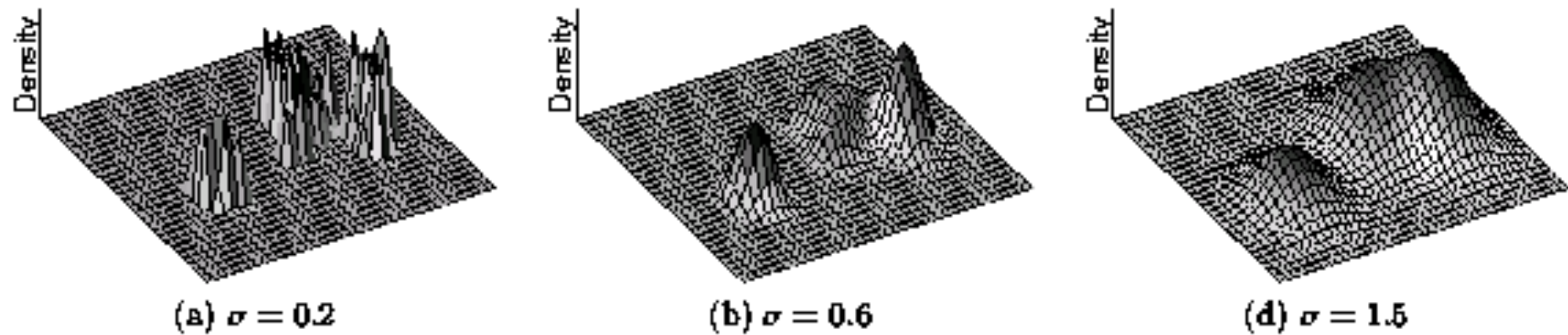


Figure 3: Example of Center-Defined Clusters for different σ

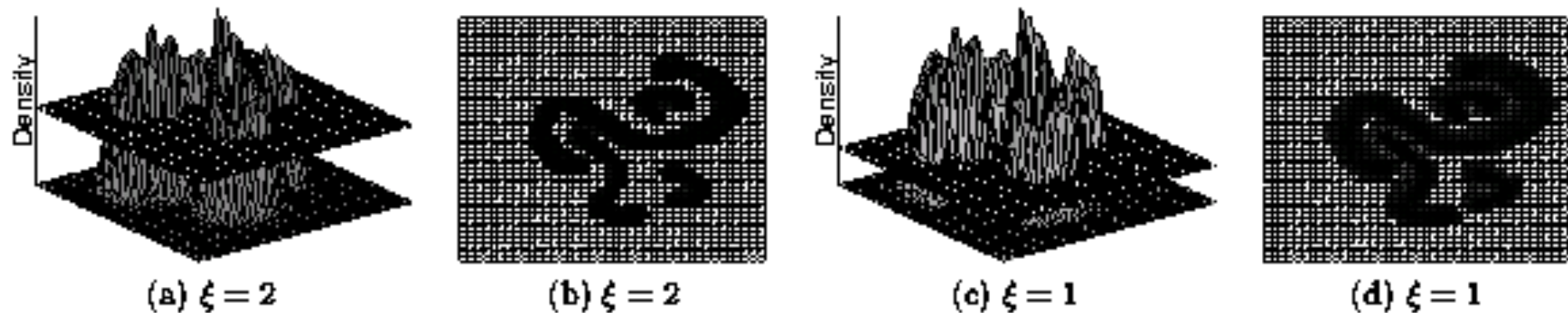



Figure 4: Example of Arbitrary-Shape Clusters for different ξ

Chapter 10. Cluster Analysis:

Khái niệm và Phương pháp cơ bản

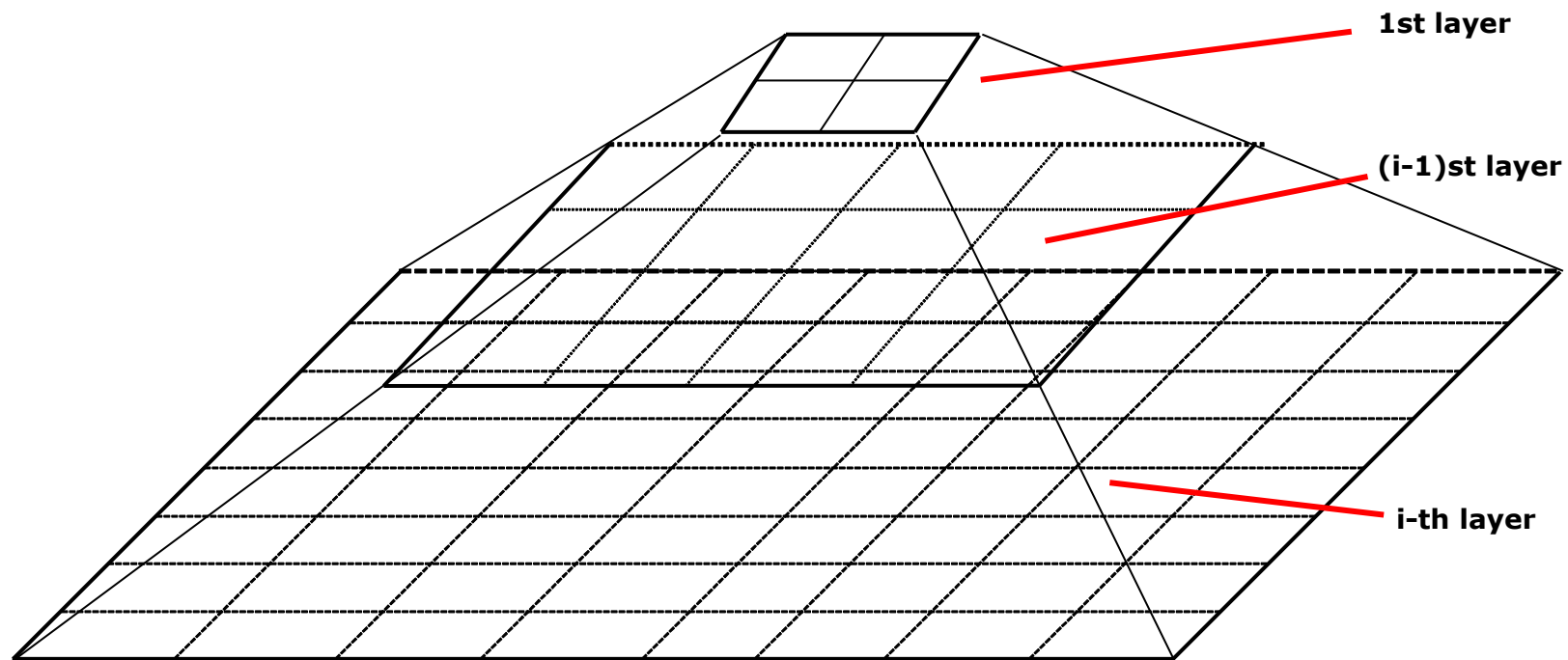
- Cluster Analysis: Basic Concepts
- Partitioning Methods
- Hierarchical Methods
- Density-Based Methods
- **Phương pháp Grid-Based** 
- Evaluation of Clustering
- Summary

Phương pháp gom cụm Grid-Based

- Using multi-resolution grid data structure
- Several interesting methods
 - **STING** (a S**T**atistical **I**Nformation Grid approach) by Wang, Yang and Muntz (1997)
 - **WaveCluster** by Sheikholeslami, Chatterjee, and Zhang (VLDB'98)
 - A multi-resolution clustering approach using wavelet method
 - **CLIQUE**: Agrawal, et al. (SIGMOD'98)
 - Both grid-based and subspace clustering

STING: A Statistical Information Grid Approach

- Wang, Yang and Muntz (VLDB'97)
- The spatial area is divided into rectangular cells
- There are several levels of cells corresponding to different levels of resolution



The STING Clustering Method

- Each cell at a high level is partitioned into a number of smaller cells in the next lower level
- Statistical info of each cell is calculated and stored beforehand and is used to answer queries
- Parameters of higher level cells can be easily calculated from parameters of lower level cell
 - *count, mean, s, min, max*
 - type of distribution—*normal, uniform, etc.*
- Use a top-down approach to answer spatial data queries
- Start from a pre-selected layer—typically with a small number of cells
- For each cell in the current level compute the confidence interval

STING Algorithm and Its Analysis

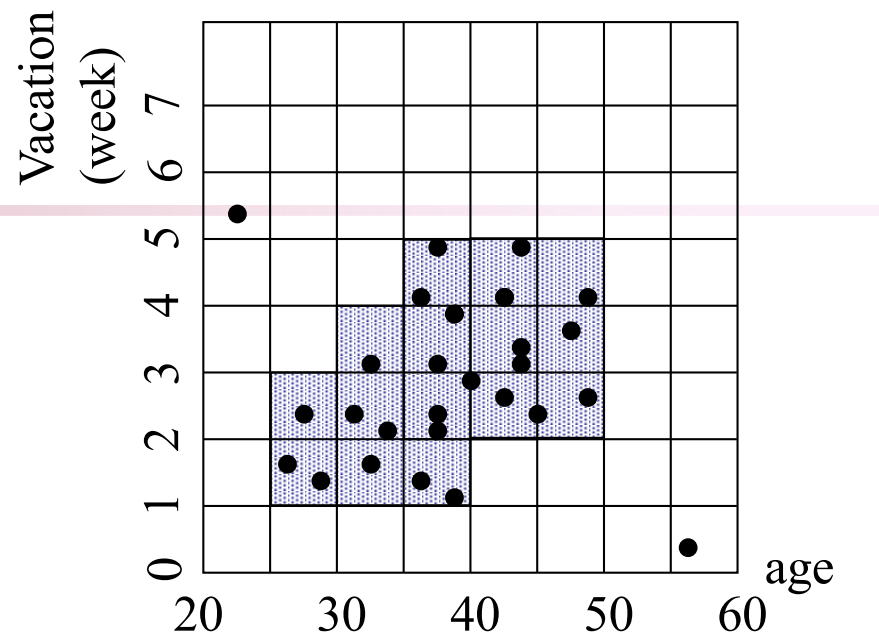
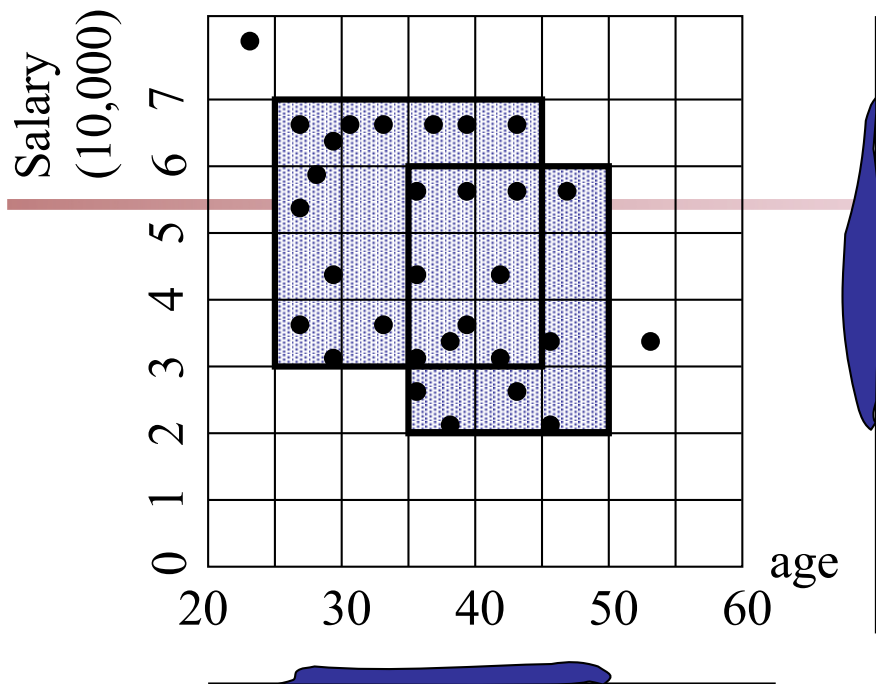
- Remove the irrelevant cells from further consideration
- When finish examining the current layer, proceed to the next lower level
- Repeat this process until the bottom layer is reached
- Advantages:
 - Query-independent, easy to parallelize, incremental update
 - $O(K)$, where K is the number of grid cells at the lowest level
- Disadvantages:
 - All the cluster boundaries are either horizontal or vertical, and no diagonal boundary is detected

CLIQUE (Clustering In QUES)

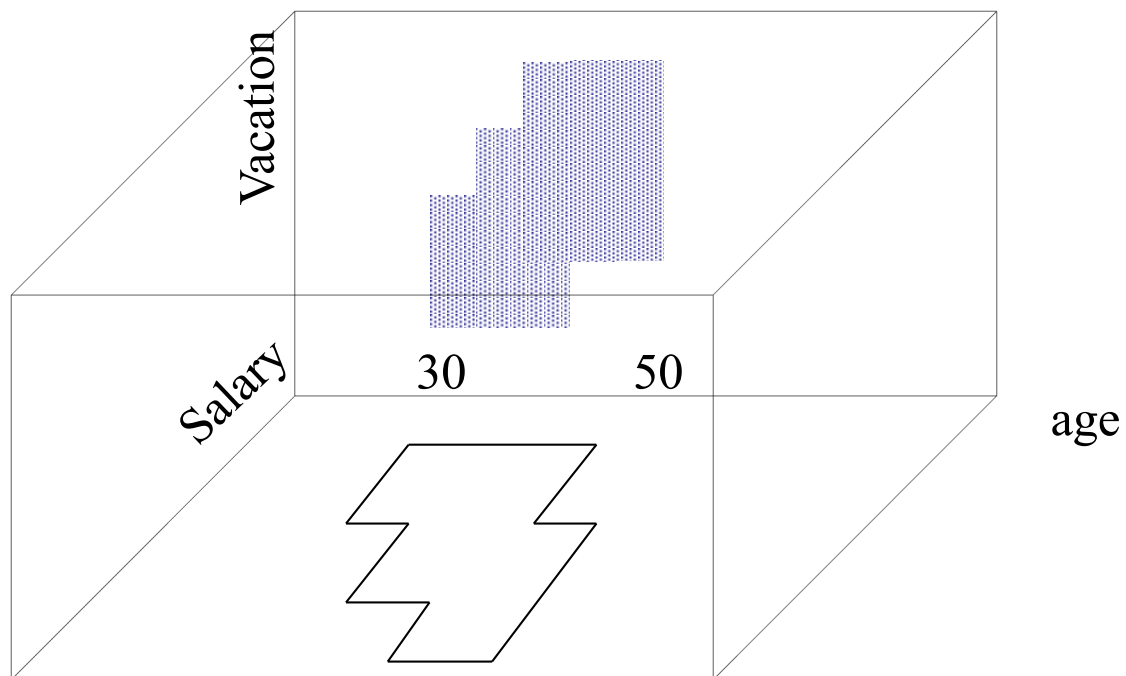
- Agrawal, Gehrke, Gunopulos, Raghavan (SIGMOD'98)
- Automatically identifying subspaces of a high dimensional data space that allow better clustering than original space
- CLIQUE can be considered as both density-based and grid-based
 - It partitions each dimension into the same number of equal length interval
 - It partitions an m-dimensional data space into non-overlapping rectangular units
 - A unit is dense if the fraction of total data points contained in the unit exceeds the input model parameter
 - A cluster is a maximal set of connected dense units within a subspace

CLIQUE: The Major Steps

- Partition the data space and find the number of points that lie inside each cell of the partition.
- Identify the subspaces that contain clusters using the Apriori principle
- Identify clusters
 - Determine dense units in all subspaces of interests
 - Determine connected dense units in all subspaces of interests.
- Generate minimal description for the clusters
 - Determine maximal regions that cover a cluster of connected dense units for each cluster
 - Determination of minimal cover for each cluster



$\tau = 3$



Strength and Weakness of *CLIQUE*


■ Strength

- *automatically* finds subspaces of the highest dimensionality such that high density clusters exist in those subspaces
- *insensitive* to the order of records in input and does not presume some canonical data distribution
- scales *linearly* with the size of input and has good scalability as the number of dimensions in the data increases

■ Weakness

- The accuracy of the clustering result may be degraded at the expense of simplicity of the method

Chapter 10. Cluster Analysis: Basic Concepts and Methods

- Cluster Analysis: Basic Concepts
- Partitioning Methods
- Hierarchical Methods
- Density-Based Methods
- Grid-Based Methods
- Evaluation of Clustering 
- Summary

Assessing Clustering Tendency

- Assess if non-random structure exists in the data by measuring the probability that the data is generated by a uniform data distribution
- Test spatial randomness by statistic test: Hopkins Static
 - Given a dataset D regarded as a sample of a random variable o , determine how far away o is from being uniformly distributed in the data space
 - Sample n points, p_1, \dots, p_n , uniformly from D . For each p_i , find its nearest neighbor in D : $x_i = \min\{\text{dist}(p_i, v)\}$ where v in D
 - Sample n points, q_1, \dots, q_n , uniformly from D . For each q_i , find its nearest neighbor in $D - \{q_i\}$: $y_i = \min\{\text{dist}(q_i, v)\}$ where v in D and $v \neq q_i$
 - Calculate the Hopkins Statistic:
$$H = \frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n x_i + \sum_{i=1}^n y_i}$$
 - If D is uniformly distributed, $\sum x_i$ and $\sum y_i$ will be close to each other and H is close to 0.5. If D is highly skewed, H is close to 0

Determine the Number of Clusters

- Empirical method
 - # of clusters $\approx \sqrt{n}/2$ for a dataset of n points
- Elbow method
 - Use the turning point in the curve of sum of within cluster variance w.r.t the # of clusters
- Cross validation method
 - Divide a given data set into m parts
 - Use $m - 1$ parts to obtain a clustering model
 - Use the remaining part to test the quality of the clustering
 - E.g., For each point in the test set, find the closest centroid, and use the sum of squared distance between all points in the test set and the closest centroids to measure how well the model fits the test set
 - For any $k > 0$, repeat it m times, compare the overall quality measure w.r.t. different k 's, and find # of clusters that fits the data the best


Measuring Clustering Quality

- Two methods: extrinsic vs. intrinsic
- Extrinsic: supervised, i.e., the ground truth is available
 - Compare a clustering against the ground truth using certain clustering quality measure
 - Ex. BCubed precision and recall metrics
- Intrinsic: unsupervised, i.e., the ground truth is unavailable
 - Evaluate the goodness of a clustering by considering how well the clusters are separated, and how compact the clusters are
 - Ex. Silhouette coefficient

Measuring Clustering Quality: Extrinsic Methods

- Clustering quality measure: $Q(C, C_g)$, for a clustering C given the ground truth C_g .
- Q is good if it satisfies the following 4 essential criteria
 - Cluster homogeneity: the purer, the better
 - Cluster completeness: should assign objects belong to the same category in the ground truth to the same cluster
 - Rag bag: putting a heterogeneous object into a pure cluster should be penalized more than putting it into a *rag bag* (i.e., “miscellaneous” or “other” category)
 - Small cluster preservation: splitting a small category into pieces is more harmful than splitting a large category into pieces

Chapter 10. Cluster Analysis: Basic Concepts and Methods

- Cluster Analysis: Basic Concepts
- Partitioning Methods
- Hierarchical Methods
- Density-Based Methods
- Grid-Based Methods
- Evaluation of Clustering
- Summary 

Summary

- **Cluster analysis** groups objects based on their **similarity** and has wide applications
- Measure of similarity can be computed for **various types of data**
- Clustering algorithms can be **categorized** into partitioning methods, hierarchical methods, density-based methods, grid-based methods, and model-based methods
- **K-means** and **K-medoids** algorithms are popular partitioning-based clustering algorithms
- **Birch** and **Chameleon** are interesting hierarchical clustering algorithms, and there are also probabilistic hierarchical clustering algorithms
- **DBSCAN**, **OPTICS**, and **DENCLU** are interesting density-based algorithms
- **STING** and **CLIQUE** are grid-based methods, where CLIQUE is also a subspace clustering algorithm
- Quality of clustering results can be evaluated in various ways

CS512-Spring 2011: An Introduction

- Coverage
 - Cluster Analysis: Chapter 11
 - Outlier Detection: Chapter 12
 - Mining Sequence Data: BK2: Chapter 8
 - Mining Graphs Data: BK2: Chapter 9
 - Social and Information Network Analysis
 - BK2: Chapter 9
 - Partial coverage: Mark Newman: “Networks: An Introduction”, Oxford U., 2010
 - Scattered coverage: Easley and Kleinberg, “Networks, Crowds, and Markets: Reasoning About a Highly Connected World”, Cambridge U., 2010
 - Recent research papers
 - Mining Data Streams: BK2: Chapter 8
- Requirements
 - One research project
 - One class presentation (15 minutes)
 - Two homeworks (no programming assignment)
 - Two midterm exams (no final exam)

References (1)

- R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan. Automatic subspace clustering of high dimensional data for data mining applications. SIGMOD'98
- M. R. Anderberg. Cluster Analysis for Applications. Academic Press, 1973.
- M. Ankerst, M. Breunig, H.-P. Kriegel, and J. Sander. Optics: Ordering points to identify the clustering structure, SIGMOD'99.
- Beil F., Ester M., Xu X.: "Frequent Term-Based Text Clustering", KDD'02
- M. M. Breunig, H.-P. Kriegel, R. Ng, J. Sander. LOF: Identifying Density-Based Local Outliers. SIGMOD 2000.
- M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases. KDD'96.
- M. Ester, H.-P. Kriegel, and X. Xu. Knowledge discovery in large spatial databases: Focusing techniques for efficient class identification. SSD'95.
- D. Fisher. Knowledge acquisition via incremental conceptual clustering. Machine Learning, 2:139-172, 1987.
- D. Gibson, J. Kleinberg, and P. Raghavan. Clustering categorical data: An approach based on dynamic systems. VLDB'98.
- V. Ganti, J. Gehrke, R. Ramakrishnan. CACTUS Clustering Categorical Data Using Summaries. KDD'99.

References (2)

- D. Gibson, J. Kleinberg, and P. Raghavan. Clustering categorical data: An approach based on dynamic systems. In Proc. VLDB'98.
- S. Guha, R. Rastogi, and K. Shim. Cure: An efficient clustering algorithm for large databases. SIGMOD'98.
- S. Guha, R. Rastogi, and K. Shim. ROCK: A robust clustering algorithm for categorical attributes. In *ICDE'99*, pp. 512-521, Sydney, Australia, March 1999.
- A. Hinneburg, D. A. Keim: An Efficient Approach to Clustering in Large Multimedia Databases with Noise. KDD'98.
- A. K. Jain and R. C. Dubes. Algorithms for Clustering Data. Printice Hall, 1988.
- G. Karypis, E.-H. Han, and V. Kumar. CHAMELEON: A Hierarchical Clustering Algorithm Using Dynamic Modeling. *COMPUTER*, 32(8): 68-75, 1999.
- L. Kaufman and P. J. Rousseeuw. Finding Groups in Data: an Introduction to Cluster Analysis. John Wiley & Sons, 1990.
- E. Knorr and R. Ng. Algorithms for mining distance-based outliers in large datasets. VLDB'98.

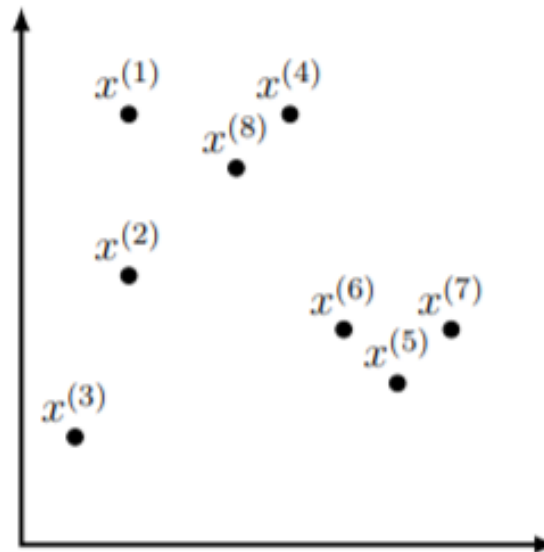
References (3)

- G. J. McLachlan and K.E. Bkasford. Mixture Models: Inference and Applications to Clustering. John Wiley and Sons, 1988.
- R. Ng and J. Han. Efficient and effective clustering method for spatial data mining. VLDB'94.
- L. Parsons, E. Haque and H. Liu, Subspace Clustering for High Dimensional Data: A Review, SIGKDD Explorations, 6(1), June 2004
- E. Schikuta. Grid clustering: An efficient hierarchical clustering method for very large data sets. Proc. 1996 Int. Conf. on Pattern Recognition
- G. Sheikholeslami, S. Chatterjee, and A. Zhang. WaveCluster: A multi-resolution clustering approach for very large spatial databases. VLDB'98.
- A. K. H. Tung, J. Han, L. V. S. Lakshmanan, and R. T. Ng. Constraint-Based Clustering in Large Databases, ICDT'01.
- A. K. H. Tung, J. Hou, and J. Han. Spatial Clustering in the Presence of Obstacles, ICDE'01
- H. Wang, W. Wang, J. Yang, and P.S. Yu. Clustering by pattern similarity in large data sets, SIGMOD'02
- W. Wang, Yang, R. Muntz, STING: A Statistical Information grid Approach to Spatial Data Mining, VLDB'97
- T. Zhang, R. Ramakrishnan, and M. Livny. BIRCH : An efficient data clustering method for very large databases. SIGMOD'96
- X. Yin, J. Han, and P. S. Yu, "LinkClus: Efficient Clustering via Heterogeneous Semantic Links", VLDB'06

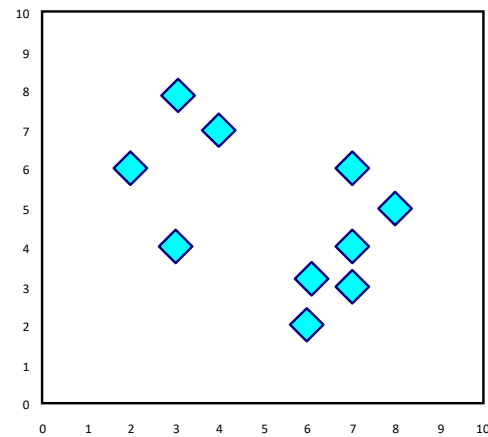
Slides unused in class

■ Exercise, k-mean clustering with $k = 3$

$$\begin{aligned}x^{(1)} &= (2, 8), & x^{(2)} &= (2, 5), & x^{(3)} &= (1, 2), & x^{(4)} &= (5, 8), \\x^{(5)} &= (7, 3), & x^{(6)} &= (6, 4), & x^{(7)} &= (8, 4), & x^{(8)} &= (4, 7).\end{aligned}$$

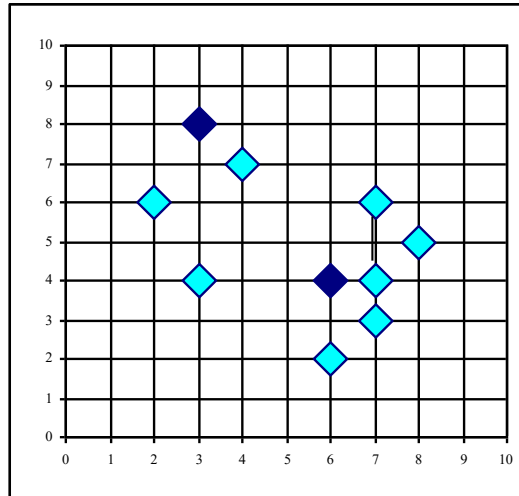


A Typical K-Medoids Algorithm (PAM)



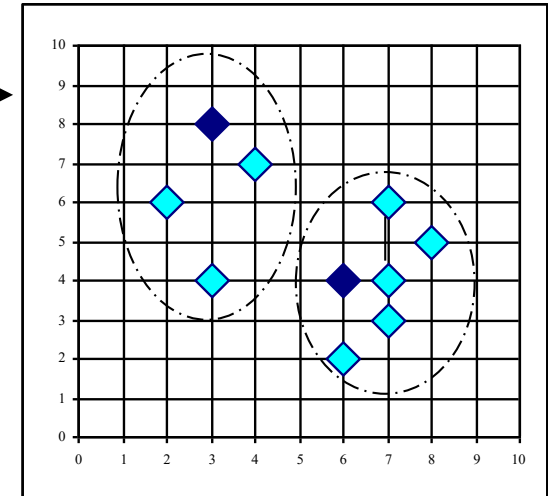
K=2

Arbitrary
choose k
object as
initial
medoids



Total Cost = 26

Assign
each
remainin
g object
to
nearest
medoids

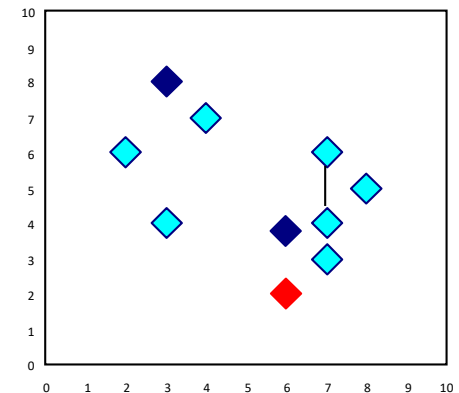
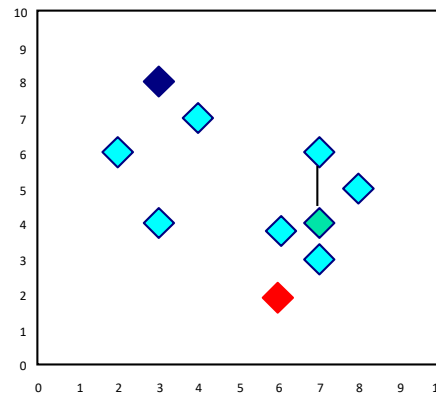


Total Cost = 20

Randomly select a
nonmedoid object, O_{random}

Compute
total cost of
swapping

Swapping O
and O_{random}
If quality is
improved.



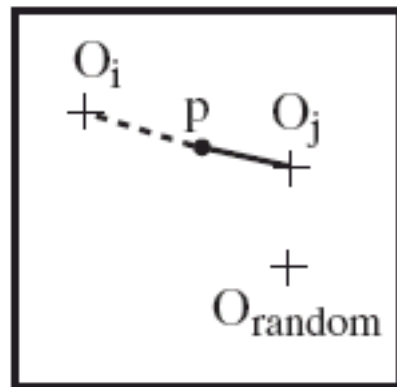
**Do loop
Until no
change**

PAM (Partitioning Around Medoids) (1987)

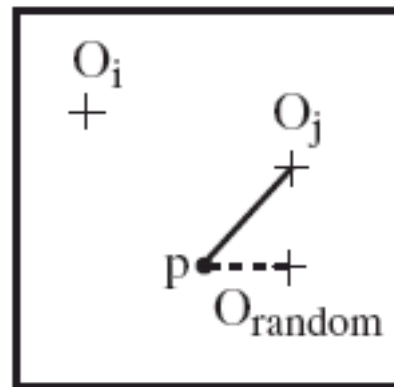
- PAM (Kaufman and Rousseeuw, 1987), built in Splus
- Use real object to represent the cluster
 - Select k representative objects arbitrarily
 - For each pair of non-selected object h and selected object i , calculate the total swapping cost TC_{ih}
 - For each pair of i and h ,
 - If $TC_{ih} < 0$, i is replaced by h
 - Then assign each non-selected object to the most similar representative object
 - repeat steps 2-3 until there is no change

PAM Clustering: Finding the Best Cluster Center

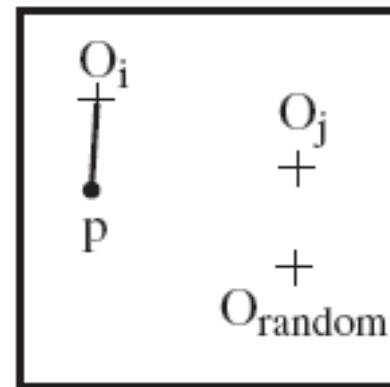
- Case 1: p currently belongs to o_j . If o_j is replaced by o_{random} as a representative object and p is the closest to one of the other representative object o_i , then p is reassigned to o_i



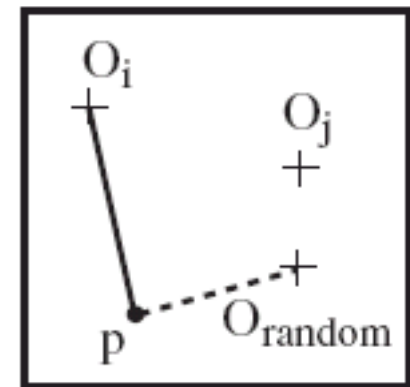
1. Reassigned to O_i



2. Reassigned to O_{random}



3. No change



4. Reassigned to O_{random}

- data object
- + cluster center
- before swapping
- after swapping

What Is the Problem with PAM?

- Pam is more robust than k-means in the presence of noise and outliers because a medoid is less influenced by outliers or other extreme values than a mean
- Pam works efficiently for small data sets but does not **scale well** for large data sets.
 - $O(k(n-k)^2)$ for each iteration

where n is # of data, k is # of clusters

➔ Sampling-based method

CLARA(Clustering LARge Applications)

CLARA (Clustering Large Applications) (1990)

- *CLARA* (Kaufmann and Rousseeuw in 1990)
 - Built in statistical analysis packages, such as SPlus
 - It draws *multiple samples* of the data set, applies *PAM* on each sample, and gives the best clustering as the output
- Strength: deals with larger data sets than *PAM*
- Weakness:
 - Efficiency depends on the sample size
 - A good clustering based on samples will not necessarily represent a good clustering of the whole data set if the sample is biased

CLARANS (“Randomized” CLARA) (1994)

- *CLARANS* (A Clustering Algorithm based on Randomized Search) (Ng and Han’94)
 - Draws sample of neighbors dynamically
 - The clustering process can be presented as searching a graph where every node is a potential solution, that is, a set of k medoids
 - If the local optimum is found, *it* starts with new randomly selected node in search for a new local optimum
- Advantages: More efficient and scalable than both *PAM* and *CLARA*
- Further improvement: Focusing techniques and spatial access structures (Ester et al.’95)

ROCK: Clustering Categorical Data

- ROCK: RObust Clustering using linkS
 - S. Guha, R. Rastogi & K. Shim, ICDE'99
- Major ideas
 - Use links to measure similarity/proximity
 - Not distance-based
- Algorithm: sampling-based clustering
 - Draw random sample
 - Cluster with links
 - Label data in disk
- Experiments
 - Congressional voting, mushroom data

Similarity Measure in ROCK

- Traditional measures for categorical data may not work well, e.g., Jaccard coefficient
- Example: Two groups (clusters) of transactions
 - C_1 . $\langle a, b, c, d, e \rangle$: $\{a, b, c\}$, $\{a, b, d\}$, $\{a, b, e\}$, $\{a, c, d\}$, $\{a, c, e\}$, $\{a, d, e\}$, $\{b, c, d\}$, $\{b, c, e\}$, $\{b, d, e\}$, $\{c, d, e\}$
 - C_2 . $\langle a, b, f, g \rangle$: $\{a, b, f\}$, $\{a, b, g\}$, $\{a, f, g\}$, $\{b, f, g\}$
- Jaccard co-efficient may lead to wrong clustering result
 - C_1 : 0.2 ($\{a, b, c\}$, $\{b, d, e\}$) to 0.5 ($\{a, b, c\}$, $\{a, b, d\}$)
 - C_1 & C_2 : could be as high as 0.5 ($\{a, b, c\}$, $\{a, b, f\}$)
- Jaccard co-efficient-based similarity function:

$$Sim(T_1, T_2) = \frac{|T_1 \cap T_2|}{|T_1 \cup T_2|}$$

- Ex. Let $T_1 = \{a, b, c\}$, $T_2 = \{c, d, e\}$

$$Sim(T_1, T_2) = \frac{|\{c\}|}{|\{a, b, c, d, e\}|} = \frac{1}{5} = 0.2$$

Link Measure in ROCK

■ Clusters

- $C_1: \langle a, b, c, d, e \rangle$: $\{a, b, c\}, \{a, b, d\}, \{a, b, e\}, \{a, c, d\}, \{a, c, e\}, \{a, d, e\}, \{b, c, d\}, \{b, c, e\}, \{b, d, e\}, \{c, d, e\}$
- $C_2: \langle a, b, f, g \rangle$: $\{a, b, f\}, \{a, b, g\}, \{a, f, g\}, \{b, f, g\}$

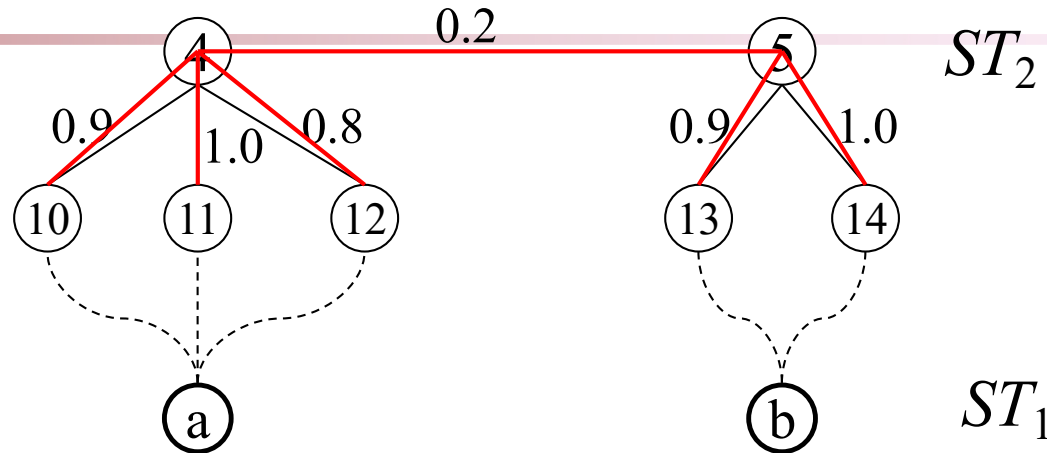
■ Neighbors

- Two transactions are neighbors if $\text{sim}(T_1, T_2) > \text{threshold}$
- Let $T_1 = \{a, b, c\}$, $T_2 = \{c, d, e\}$, $T_3 = \{a, b, f\}$
 - T_1 connected to: $\{a, b, d\}, \{a, b, e\}, \{a, c, d\}, \{a, c, e\}, \{b, c, d\}, \{b, c, e\}, \{a, b, f\}, \{a, b, g\}$
 - T_2 connected to: $\{a, c, d\}, \{a, c, e\}, \{a, d, e\}, \{b, c, e\}, \{b, d, e\}, \{b, c, d\}$
 - T_3 connected to: $\{a, b, c\}, \{a, b, d\}, \{a, b, e\}, \{a, b, g\}, \{a, f, g\}, \{b, f, g\}$

■ Link Similarity

- Link similarity between two transactions is the # of common neighbors
- $\text{link}(T_1, T_2) = 4$, *since they have 4 common neighbors*
 - $\{a, c, d\}, \{a, c, e\}, \{b, c, d\}, \{b, c, e\}$
- $\text{link}(T_1, T_3) = 3$, *since they have 3 common neighbors*
 - $\{a, b, d\}, \{a, b, e\}, \{a, b, g\}$

Aggregation-Based Similarity Computation



For each node $n_k \in \{n_{10}, n_{11}, n_{12}\}$ and $n_l \in \{n_{13}, n_{14}\}$, their path-based similarity $sim_p(n_k, n_l) = s(n_k, n_4) \cdot s(n_4, n_5) \cdot s(n_5, n_l)$.

$$sim(n_a, n_b) = \frac{\sum_{k=10}^{12} s(n_k, n_4)}{3} \cdot s(n_4, n_5) \cdot \frac{\sum_{l=13}^{14} s(n_l, n_5)}{2} = 0.171$$

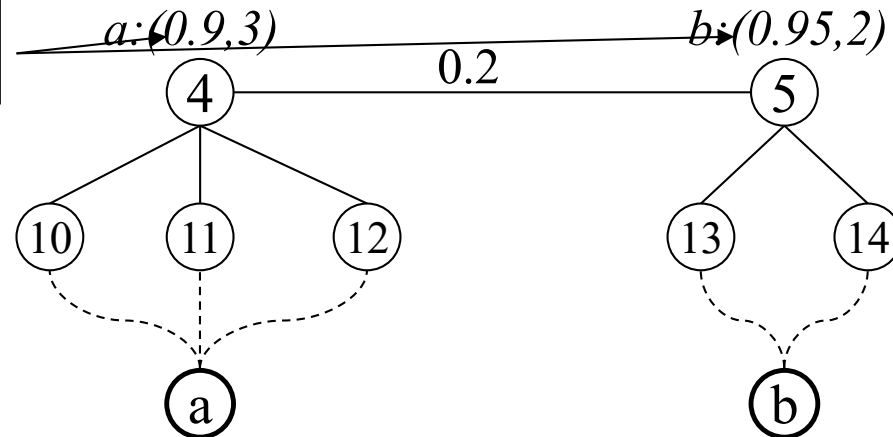
takes $O(3+2)$ time

After aggregation, we reduce quadratic time computation to linear time computation.

Computing Similarity with Aggregation

Average similarity
and total weight

$sim(n_a, n_b)$ can be computed
from aggregated similarities




$$sim(n_a, n_b) = avg_sim(n_a, n_4) \times s(n_4, n_5) \times avg_sim(n_b, n_5)$$

$$= 0.9 \times 0.2 \times 0.95 = 0.171$$

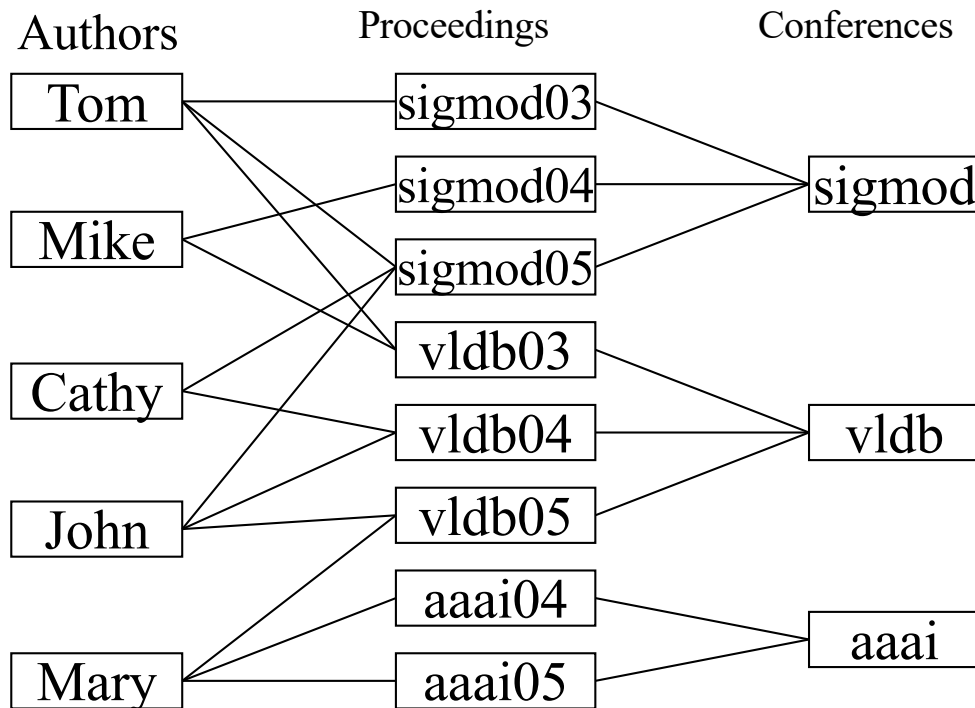
To compute $sim(n_a, n_b)$:

- Find all pairs of sibling nodes n_i and n_j , so that n_a linked with n_i and n_b with n_j .
- Calculate similarity (and weight) between n_a and n_b w.r.t. n_i and n_j .
- Calculate weighted average similarity between n_a and n_b w.r.t. all such pairs.

Chapter 10. Cluster Analysis: Basic Concepts and Methods

- Cluster Analysis: Basic Concepts
- Overview of Clustering Methods
- Partitioning Methods
- Hierarchical Methods
- Density-Based Methods 
- Grid-Based Methods
- Summary

Link-Based Clustering: Calculate Similarities Based On Links



- The similarity between two objects x and y is defined as the average similarity between objects linked with x and those with y :

$$\text{sim}(a, b) = \frac{C}{|I(a)||I(b)|} \sum_{i=1}^{|I(a)|} \sum_{j=1}^{|I(b)|} \text{sim}(I_i(a), I_j(b))$$

- Issue: Expensive to compute:
 - For a dataset of N objects and M links, it takes $O(N^2)$ space and $O(M^2)$ time to compute all similarities.

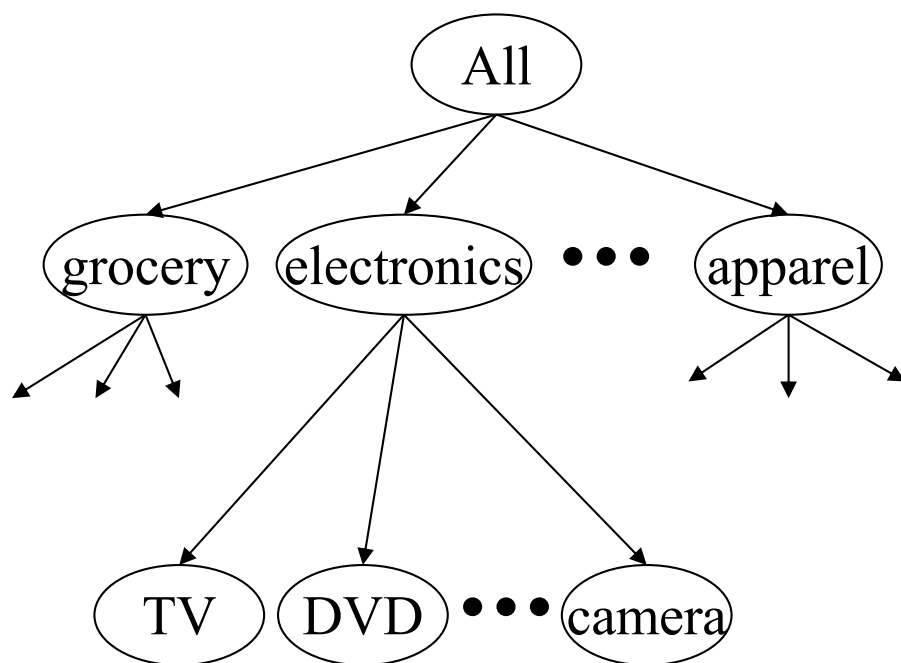
Jeh & Widom, KDD'2002: *SimRank*

Two objects are similar if they are linked with the same or similar objects

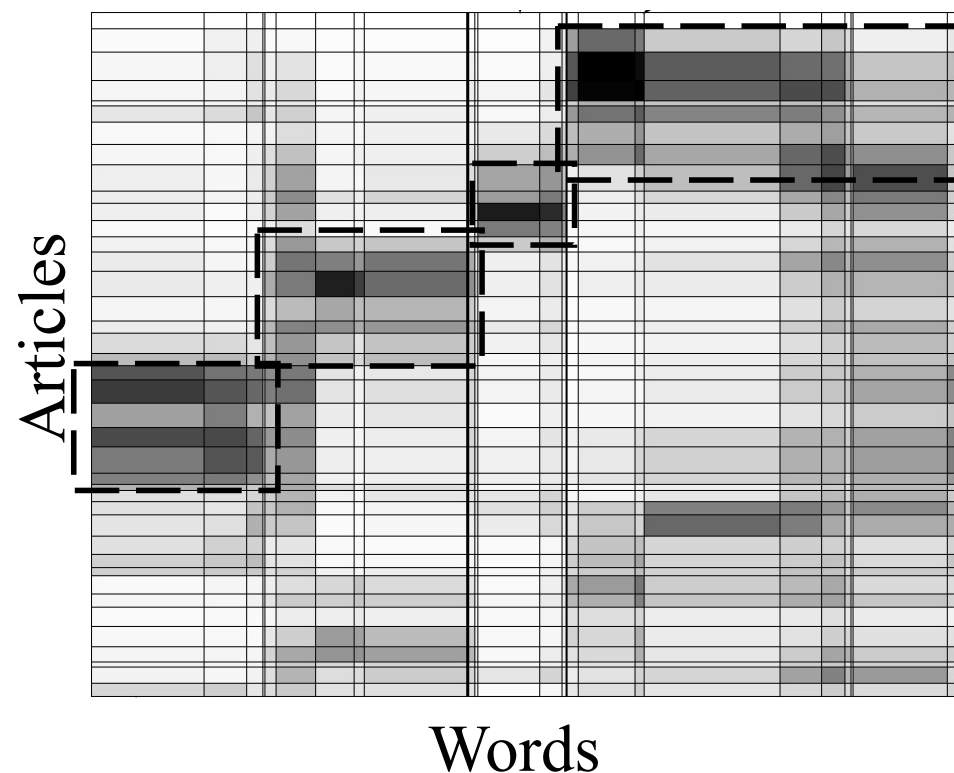
Observation 1: Hierarchical Structures

- Hierarchical structures often exist naturally among objects (e.g., taxonomy of animals)

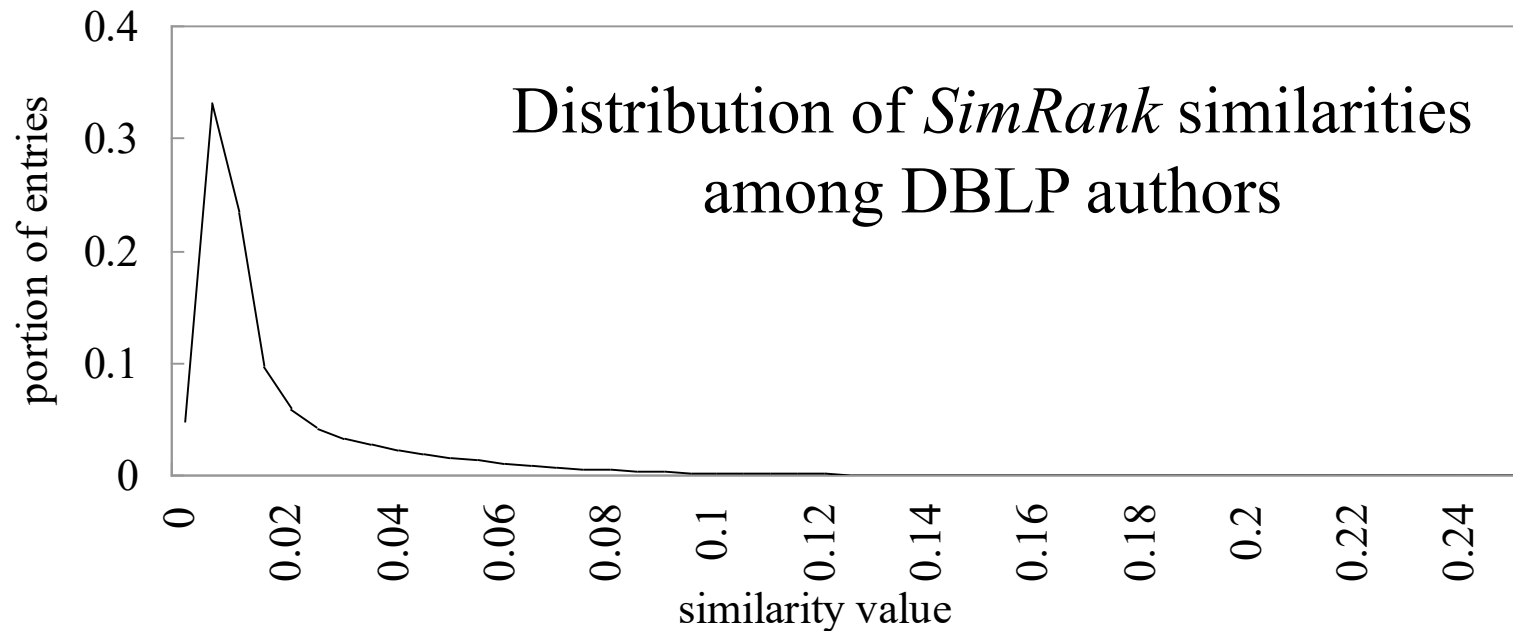
A hierarchical structure of products in Walmart



Relationships between articles and words (Chakrabarti, Papadimitriou, Modha, Faloutsos, 2004)

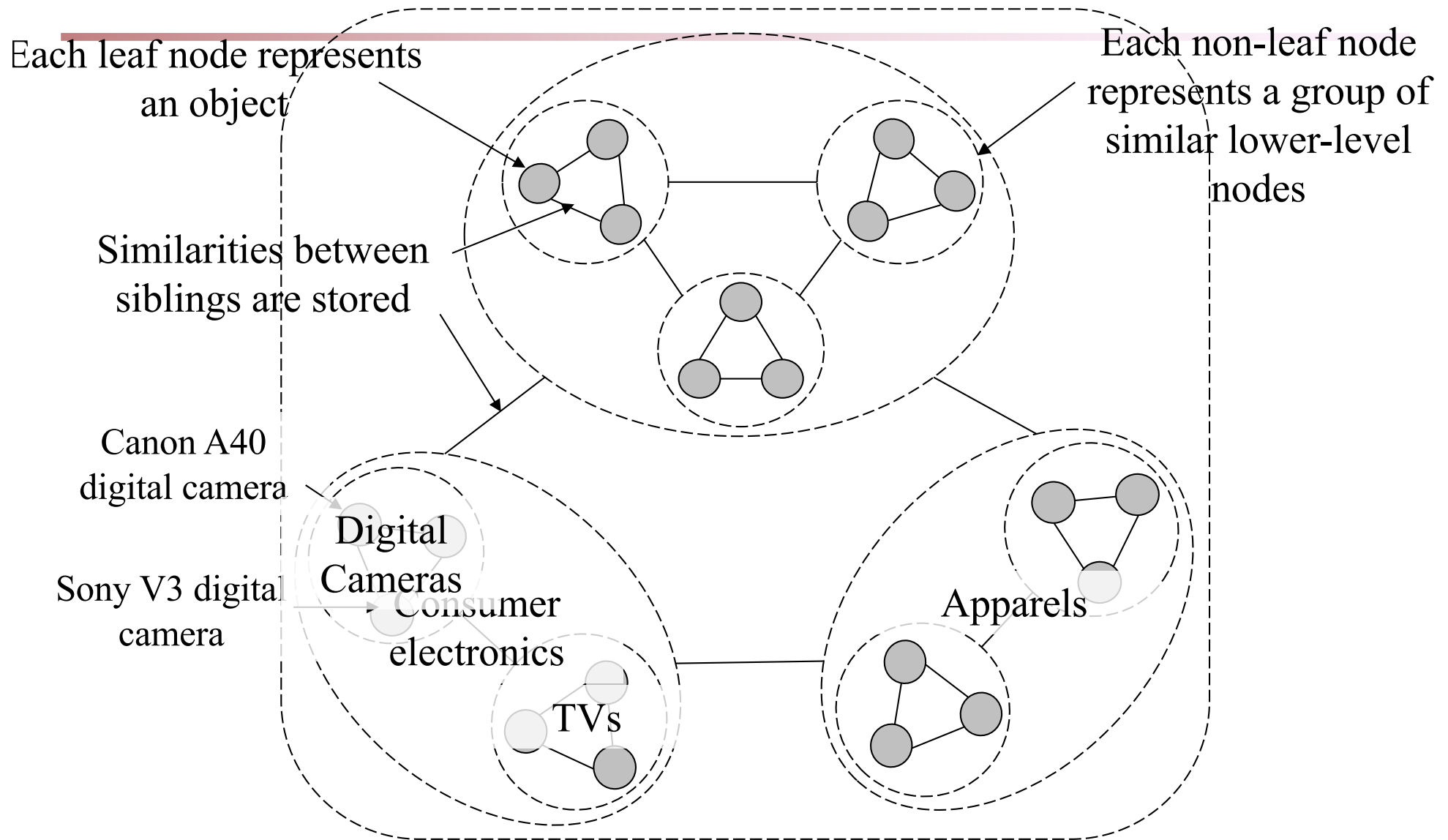


Observation 2: Distribution of Similarity

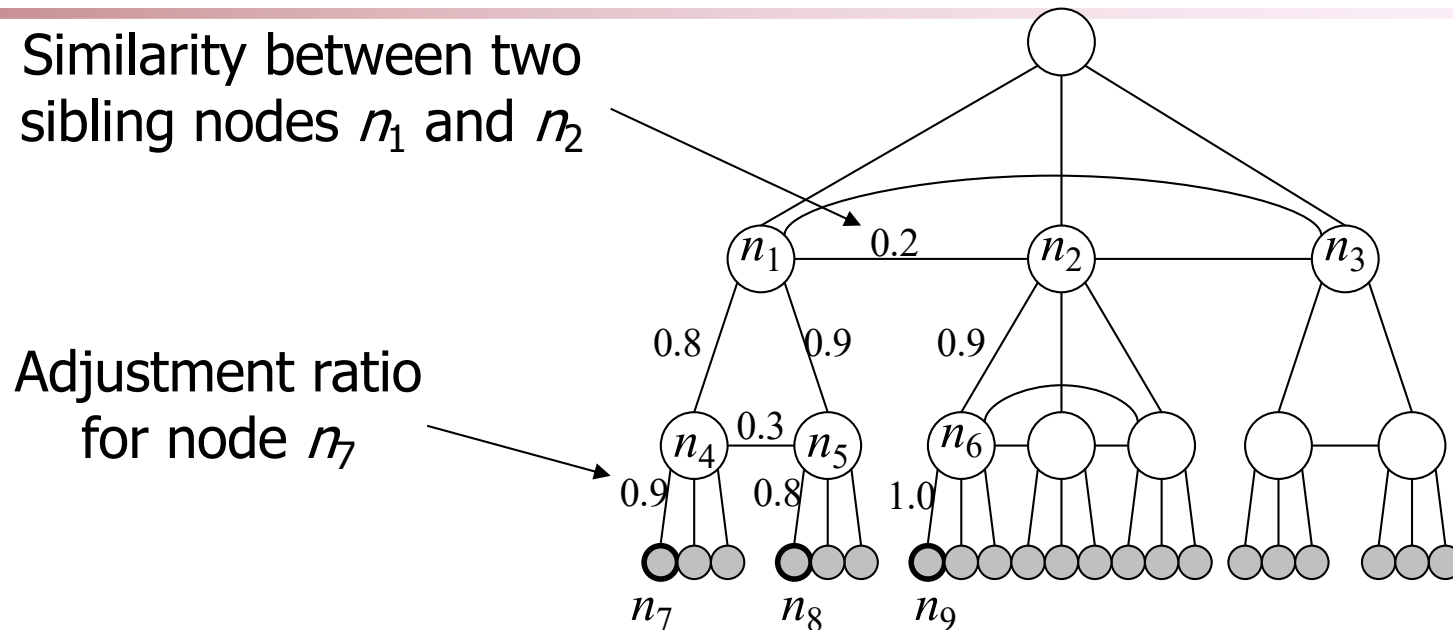


- Power law distribution exists in similarities
 - 56% of similarity entries are in $[0.005, 0.015]$
 - 1.4% of similarity entries are larger than 0.1
 - Can we design a data structure that stores the significant similarities and compresses insignificant ones?

A Novel Data Structure: SimTree



Similarity Defined by SimTree



- Path-based node similarity
 - $sim_p(n_7, n_8) = s(n_7, n_4) \times s(n_4, n_5) \times s(n_5, n_8)$
- Similarity between two nodes is the average similarity between objects linked with them in other SimTrees
- Adjust/ ratio for $x = \frac{\text{Average similarity between } x \text{ and all other nodes}}{\text{Average similarity between } x\text{'s parent and all other nodes}}$

LinkClus: Efficient Clustering via Heterogeneous Semantic Links

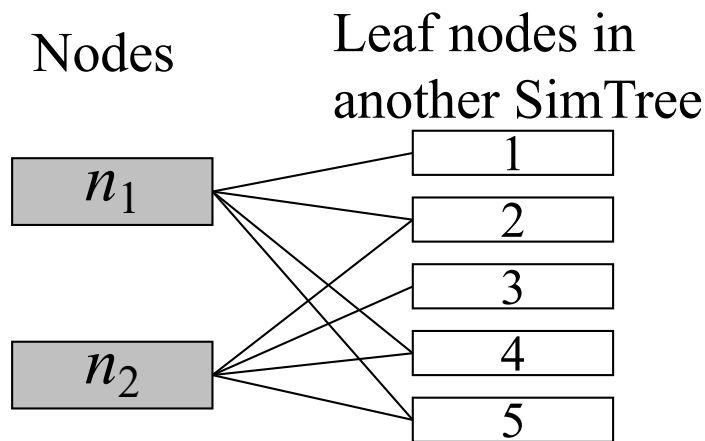
Method

- Initialize a SimTree for objects of each type
- Repeat until stable
 - For each SimTree, update the similarities between its nodes using similarities in other SimTrees
 - Similarity between two nodes x and y is the average similarity between objects linked with them
 - Adjust the structure of each SimTree
 - Assign each node to the parent node that it is most similar to

For details: X. Yin, J. Han, and P. S. Yu, “LinkClus: Efficient Clustering via Heterogeneous Semantic Links”, VLDB'06

Initialization of SimTrees

- Initializing a SimTree
 - Repeatedly find groups of tightly related nodes, which are merged into a higher-level node
- Tightness of a group of nodes
 - For a group of nodes $\{n_1, \dots, n_k\}$, its tightness is defined as the number of leaf nodes in other SimTrees that are connected to all of $\{n_1, \dots, n_k\}$

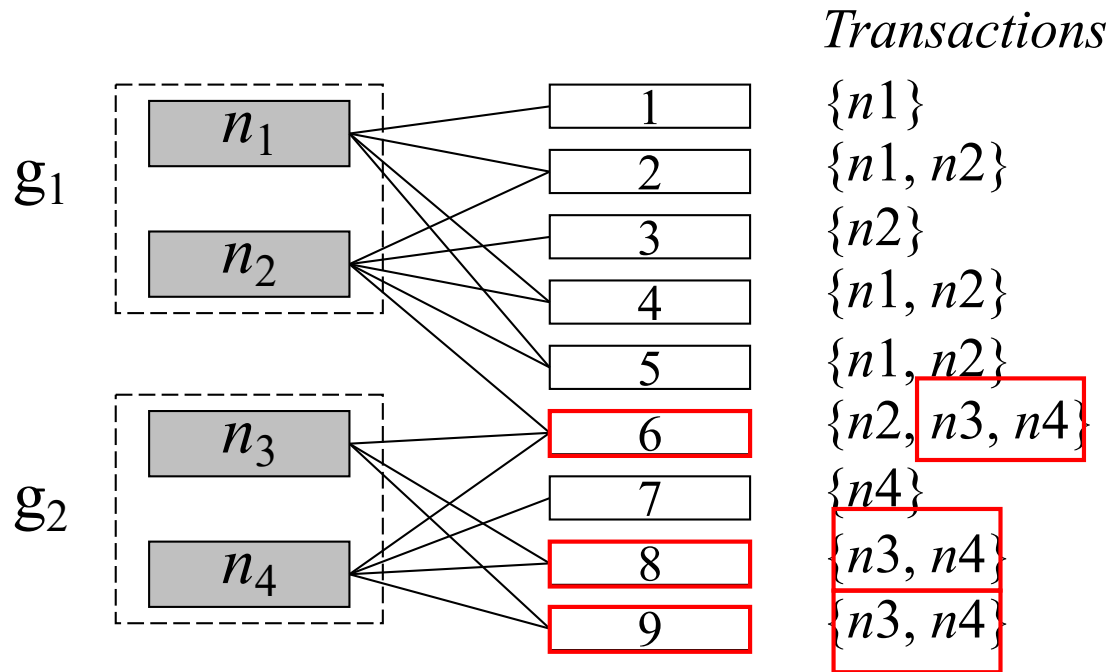


The tightness of $\{n_1, n_2\}$ is 3

Finding Tight Groups by Freq. Pattern Mining

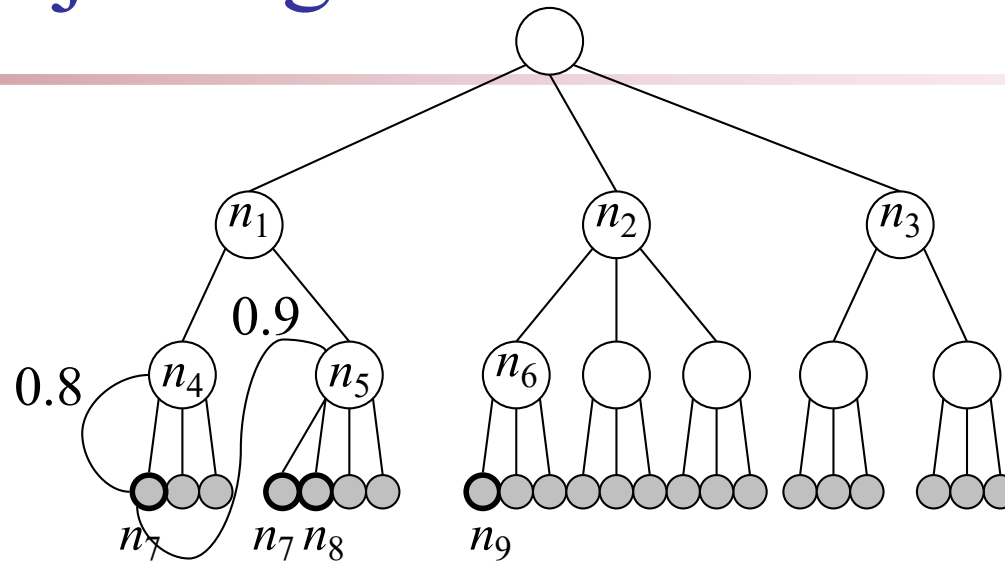
- Finding tight groups \longrightarrow Frequent pattern mining
Reduced to

The tightness of a group of nodes is the support of a frequent pattern



- Procedure of initializing a tree
 - Start from leaf nodes (level-0)
 - At each level l , find non-overlapping groups of similar nodes with frequent pattern mining

Adjusting SimTree Structures



- After similarity changes, the tree structure also needs to be changed
 - If a node is more similar to its parent's sibling, then move it to be a child of that sibling
 - Try to move each node to its parent's sibling that it is most similar to, under the constraint that each parent node can have at most c children

Complexity

For two types of objects, N in each, and M linkages between them.

| | Time | Space |
|---------------------------|------------------|----------|
| Updating similarities | $O(M(\log N)^2)$ | $O(M+N)$ |
| Adjusting tree structures | $O(N)$ | $O(N)$ |
| | | |
| LinkClus | $O(M(\log N)^2)$ | $O(M+N)$ |
| SimRank | $O(M^2)$ | $O(N^2)$ |

Experiment: Email Dataset

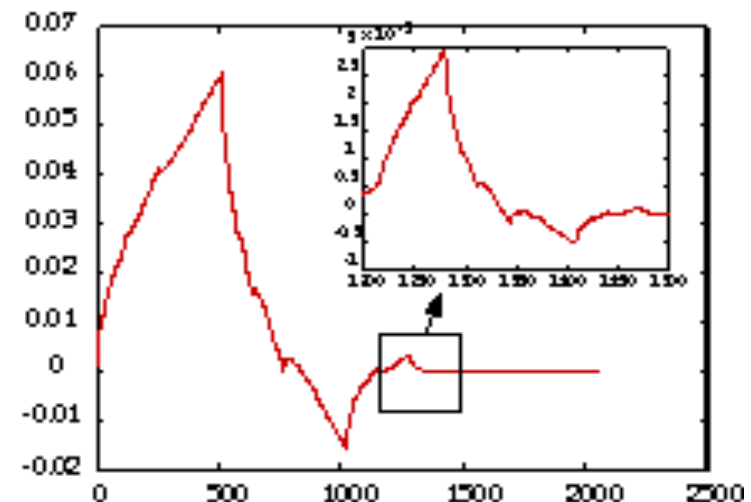
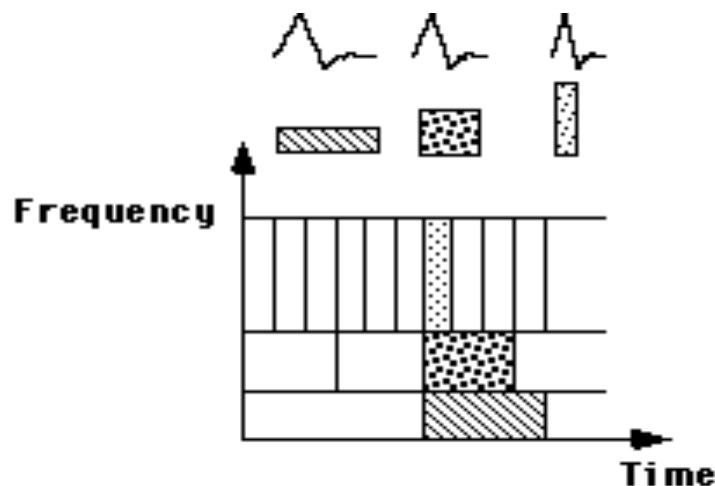
- F. Nielsen. Email dataset.
www.imm.dtu.dk/~rem/data/Email-1431.zip
- 370 emails on conferences, 272 on jobs, and 789 spam emails
- Accuracy: measured by manually labeled data
- Accuracy of clustering: % of pairs of objects in the same cluster that share common label

| Approach | Accuracy | time (s) |
|-----------|----------|----------|
| LinkClus | 0.8026 | 1579.6 |
| SimRank | 0.7965 | 39160 |
| ReCom | 0.5711 | 74.6 |
| F-SimRank | 0.3688 | 479.7 |
| CLARANS | 0.4768 | 8.55 |

- Approaches compared:
 - SimRank (Jeh & Widom, KDD 2002): Computing pair-wise similarities
 - SimRank with FingerPrints (F-SimRank): Fogaras & R'acz, WWW 2005
 - pre-computes a large sample of random paths from each object and uses samples of two objects to estimate SimRank similarity
 - ReCom (Wang et al. SIGIR 2003)
 - Iteratively clustering objects using cluster labels of linked objects

WaveCluster: Clustering by Wavelet Analysis (1998)

- Sheikholeslami, Chatterjee, and Zhang (VLDB'98)
- A multi-resolution clustering approach which applies wavelet transform to the feature space; both grid-based and density-based
- Wavelet transform: A signal processing technique that decomposes a signal into different frequency sub-band
 - Data are transformed to preserve relative distance between objects at different levels of resolution
 - Allows natural clusters to become more distinguishable



The WaveCluster Algorithm

- How to apply wavelet transform to find clusters
 - Summarizes the data by imposing a multidimensional grid structure onto data space
 - These multidimensional spatial data objects are represented in a n-dimensional feature space
 - Apply wavelet transform on feature space to find the dense regions in the feature space
 - Apply wavelet transform multiple times which result in clusters at different scales from fine to coarse
- Major features:
 - Complexity $O(N)$
 - Detect arbitrary shaped clusters at different scales
 - Not sensitive to noise, not sensitive to input order
 - Only applicable to low dimensional data

Quantization & Transformation

- Quantize data into m-D grid structure
wavelet transform
 - a) scale 1: high resolution
 - b) scale 2: medium resolution
 - c) scale 3: low resolution

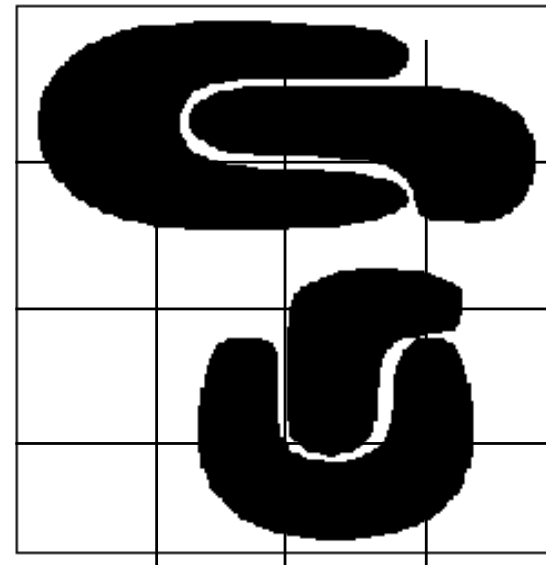
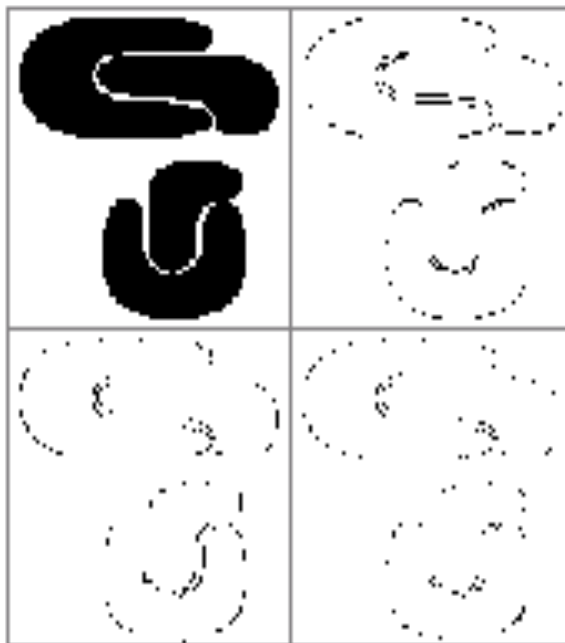


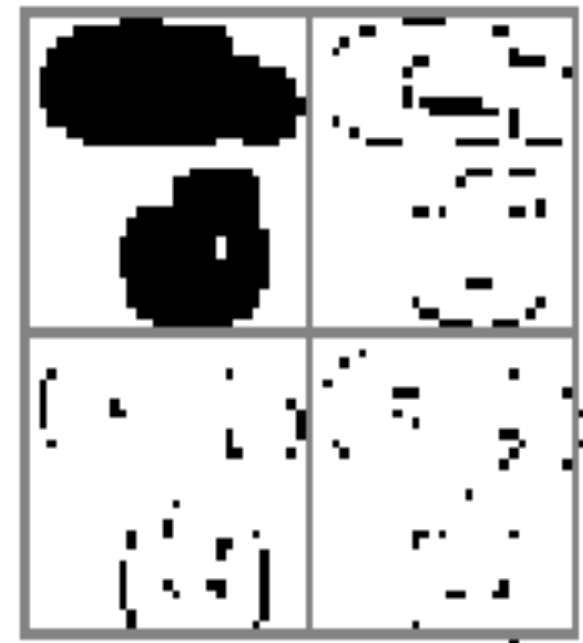
Figure 1: A sample 2-dimensional feature space.



a)



b)



c)