Clustering Categorical Data: The ROCK Algorithm

- ROCK: RObust Clustering using linKs, ICDE'99
- Major ideas
  - Use the notion of *links* to measure similarity/proximity
    - Not distance-based

Similarity Measure in ROCK

- Traditional measures for categorical data may not work well, e.g., Jaccard coefficient
- *Jaccard coefficient*-based similarity function:

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

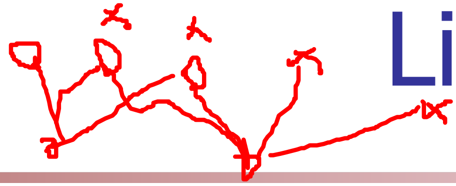
  - Ex. Let T_1 = {a, b, c}, T_2 = {c, d, e}

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

Similarity Measure in ROCK

- Example: Two groups (clusters) of transactions
  - C_1. <a, b, c, d, e>: {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. <a, b, f, g>: {a, b, f}, {a, b, g}, {a, f, g}, {b, f, g}
- Jaccard coefficient may lead to a 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})

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



Link Measure in ROCK

- Links: # of common *neighbors* (threshold = 0.5 in jC)
  - C_1 <a, b, c, d, e>: **{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 <a, b, f, g>: **{a, b, f}**, {a, b, g}, {a, f, g}, {b, f, g}
- Let T_1 = {**a**, **b**, **c**}, T_2 = {**c**, d, e}, T_3 = {**a**, **b**, f}
  - link(T_1, T_2) = 4, *since they have 4 common neighbors*
    - {a, c, d}, {a, c, e}, {b, c, d}, {b, c, e}
  - link(T_1, T_3) = 3, *since they have 3 common neighbors*
    - {a, b, d}, {a, b, e}, {a, b, g}
- Thus, link is a better measure than Jaccard coefficient

CHAMELEON: Hierarchical Clustering Using Dynamic Modeling (1999)

- CHAMELEON: by G. Karypis, E.H. Han, and V. Kumar'99
- 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 internal closeness of items within the clusters

CHAMELEON: Hierarchical Clustering Using Dynamic Modeling (1999)

- Draw a k-nearest neighbor graph first
  - Node: object, edge: k-nearest neighbor's link, weight: similarity
- A two-phase algorithm
  - Use a graph partitioning algorithm:
    - Cluster objects into a large number of relatively small sub-clusters
  - Use an agglomerative hierarchical clustering algorithm:
    - Find the genuine clusters by repeatedly combining these sub-clusters

CHAMELEON: Hierarchical Clustering Using Dynamic Modeling (1999)

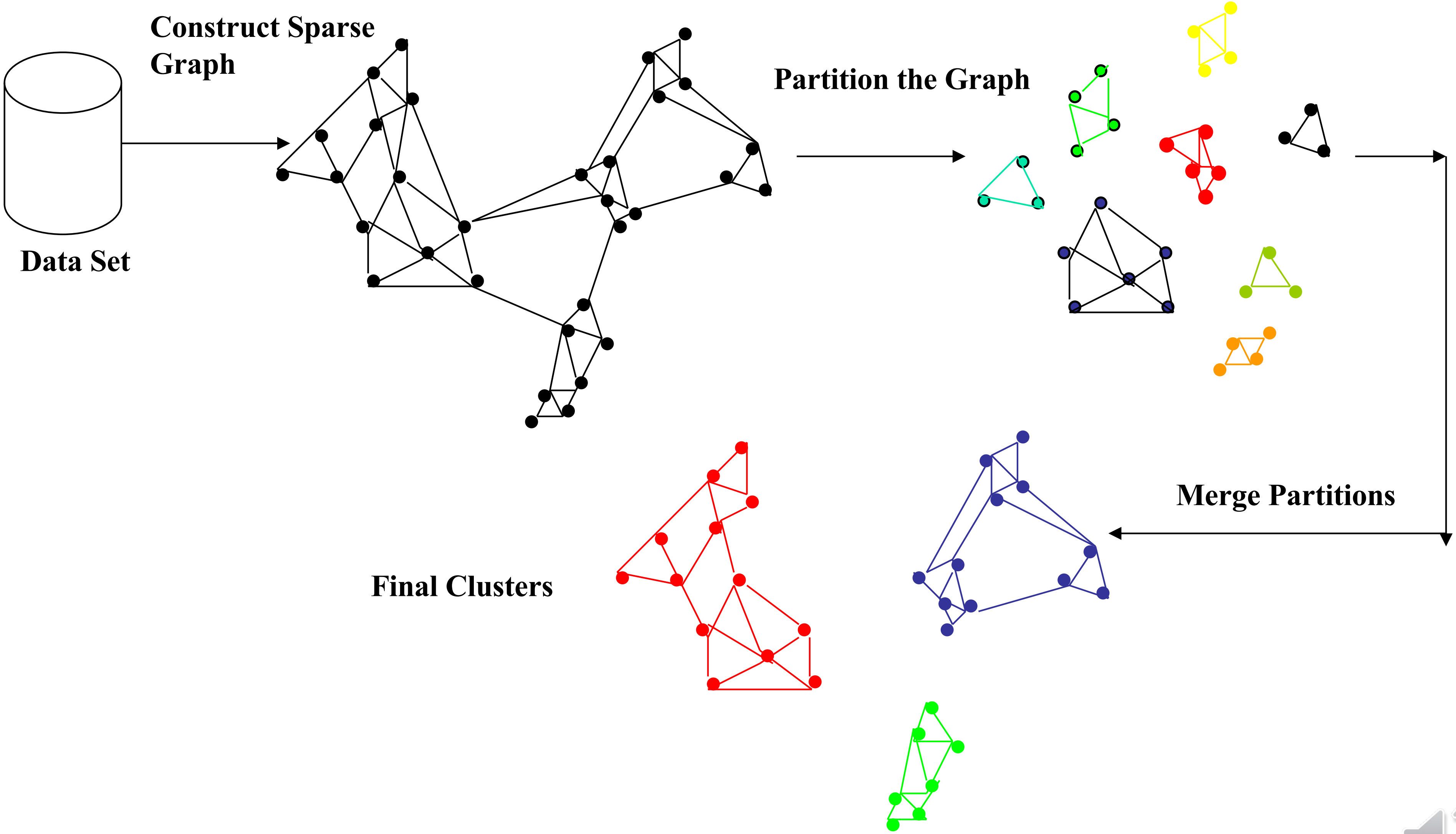
- Partitioning
  - To minimize the edge cut (**METIS**)
    - Tries to split a graph into two subgraphs of nearly equal sizes
- Relative interconnectivity

$$RI(C_i, C_j) = \frac{|EC_{\{C_i, C_j\}}|}{\frac{1}{2}(|EC_{C_i}| + |EC_{C_j}|)},$$

- Relative closeness

$$RC(C_i, C_j) = \frac{\bar{S}_{EC_{\{C_i, C_j\}}}}{\frac{|C_i|}{|C_i|+|C_j|}\bar{S}_{EC_{C_i}} + \frac{|C_j|}{|C_i|+|C_j|}\bar{S}_{EC_{C_j}}},$$

Overall Framework of CHAMELEON



CHAMELEON (Clustering Complex Objects)

Chapter 7. Cluster Analysis

1. What is Cluster Analysis?
2. Types of Data in Cluster Analysis
3. A Categorization of Major Clustering Methods
4. Partitioning Methods
5. Hierarchical Methods
6. Density-Based Methods
7. Outlier Analysis
8. Summary

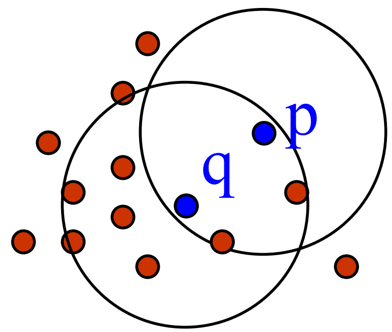
Density-Based Clustering Methods

- Clustering based on density (local cluster criterion), such as density-connected points, rather than just a distance
- Major features:
  - Discover clusters of arbitrary shape
  - Handle noise
  - One scan, thus being efficient
  - Need density parameters as termination condition
- Several interesting studies:
  - DBSCAN: Ester, et al. (KDD'96)
  - OPTICS: Ankerst, et al (SIGMOD'99).
  - CLIQUE: Agrawal, et al. (SIGMOD'98) (more grid-based)

Density-Based Clustering: Basic Concepts

- Two parameters*:*
  - *Eps*: Maximum radius of the neighborhood
  - *MinPts*: Minimum number of points in an Eps-neighborhood of a given point
- $N_{Eps}(p)$: $\{q \text{ belongs to } D \mid dist(p,q) <= Eps\}$
- Directly density-reachable: A point *p* is directly density-reachable from a point *q* w.r.t. *Eps* and *MinPts* if
  - *p* belongs to $N_{Eps}(q)$
  - core point condition:

    $|N_{Eps}(q)| >= MinPts$
  - Note: *Not symmetric*

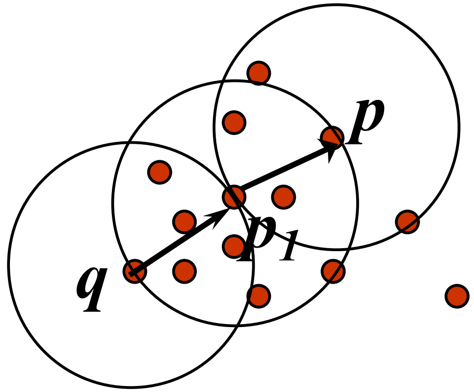


MinPts = 5

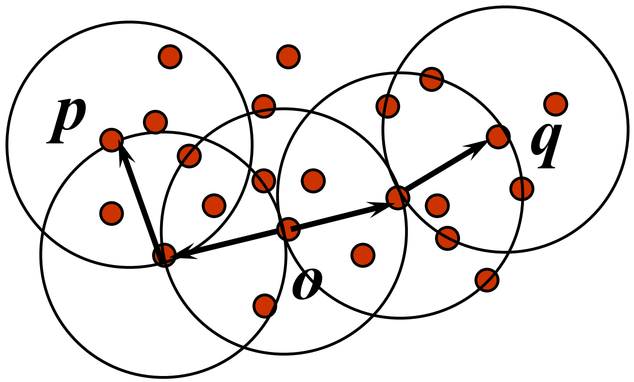
Eps = 1 cm

Density-Reachable and Density-Connected

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

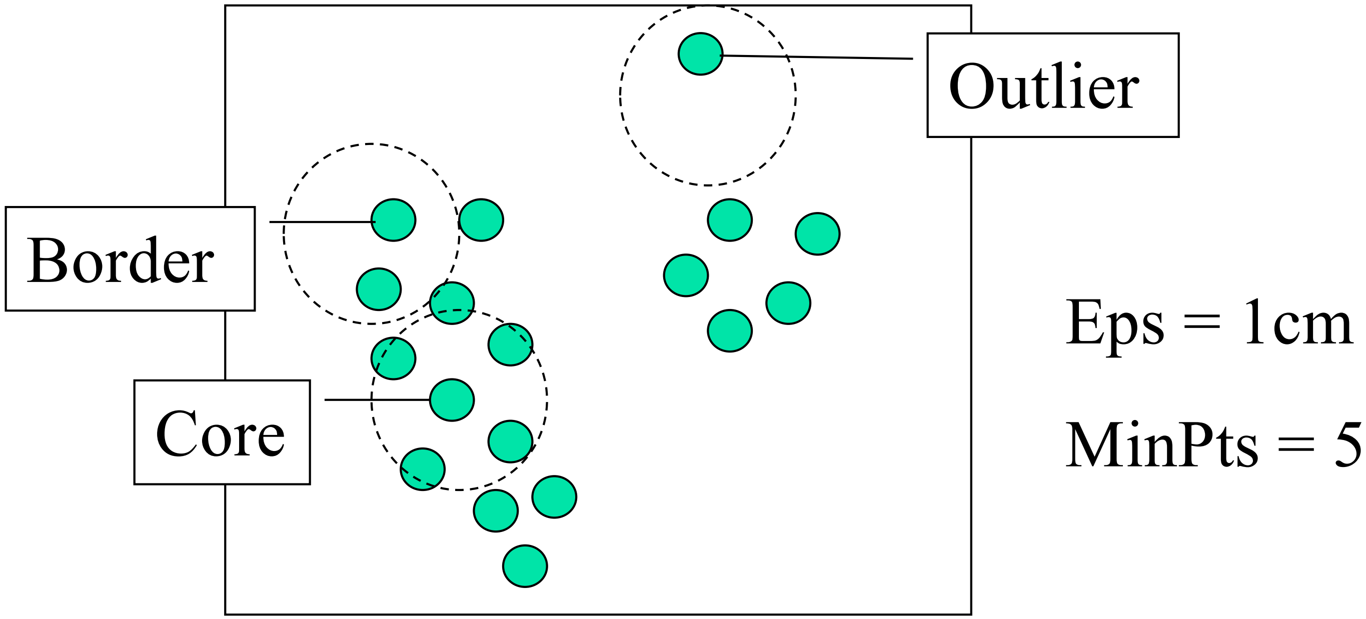


- Density-connected
  - A point *p* is density-connected to a point *q* w.r.t. *Eps* and *MinPts* if there is a point *o* such that both, *p* and *q* are density-reachable from *o* w.r.t. *Eps* and *MinPts*



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 an *arbitrary shape* in spatial databases with noise



DBSCAN: The Algorithm

- 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: Sensitive to Parameters

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

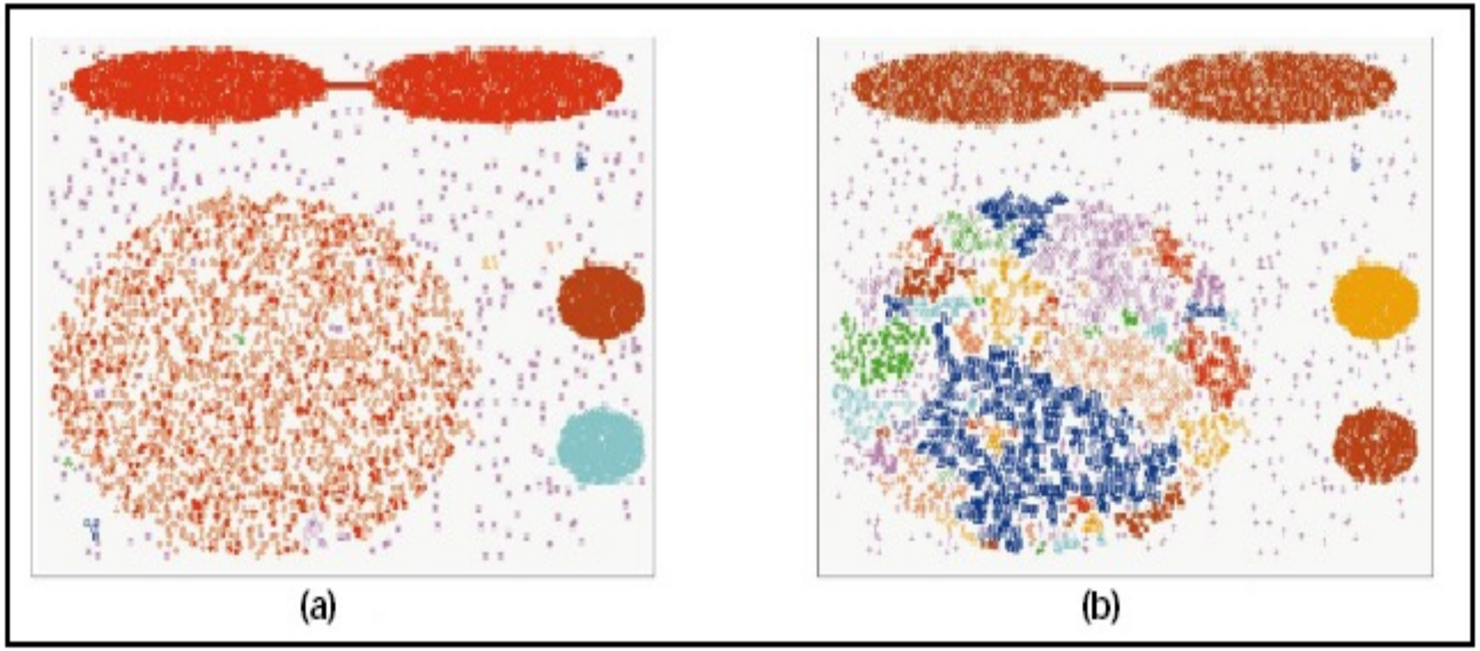
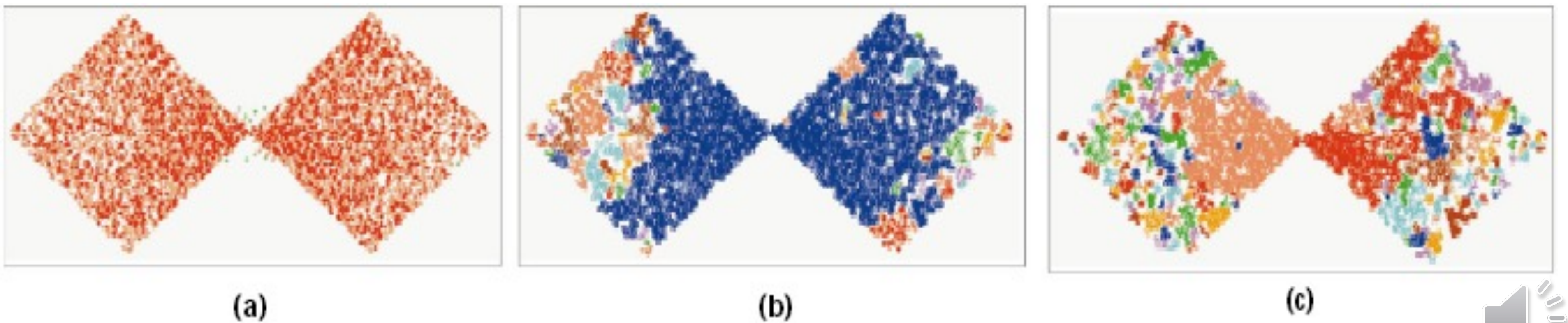
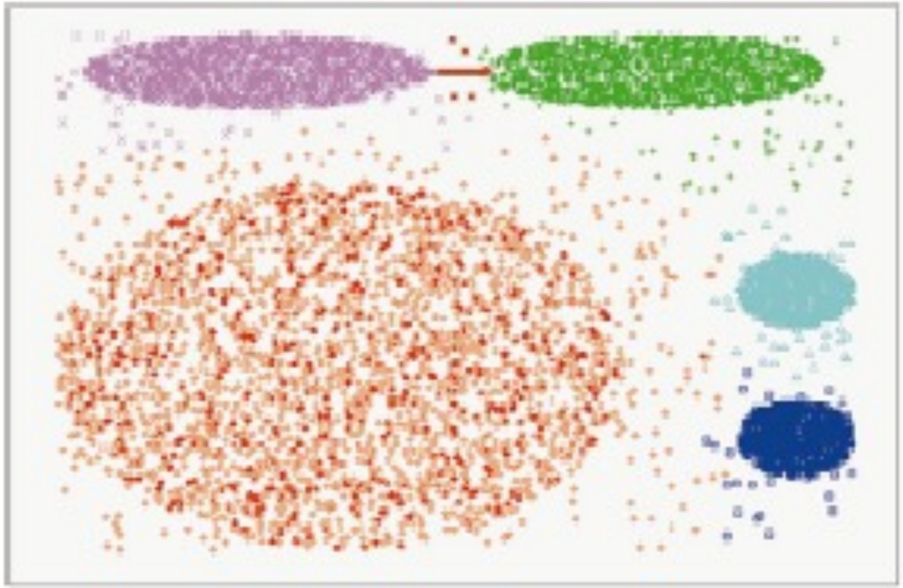
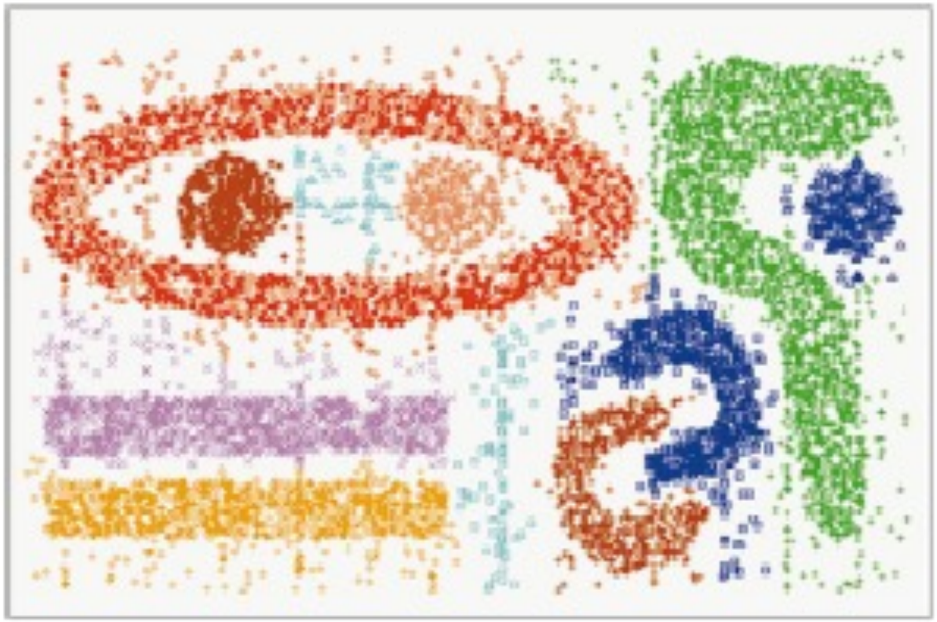


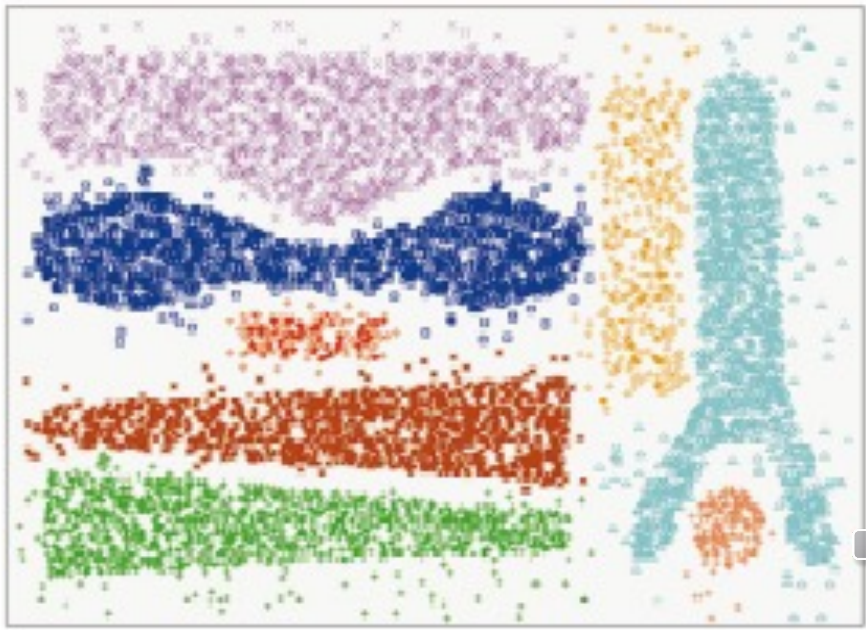
Figure 9. DBScan results for DS2 with MinPts at 4 and Eps at (a) 5.0, (b) 3.5, and (c) 3.0.



CHAMELEON (Clustering Complex Objects)







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 objects in the database w.r.t. its density-based clustering structure
  - This cluster-ordering contains the information equivalent to different density-based clustering structure corresponding to a broad range of parameter settings (*Eps*)
  - Good for both automatic and interactive cluster analysis, including finding intrinsic clustering structure
  - Can be represented graphically or using visualization techniques

OPTICS: A Cluster-Ordering Method (1999)

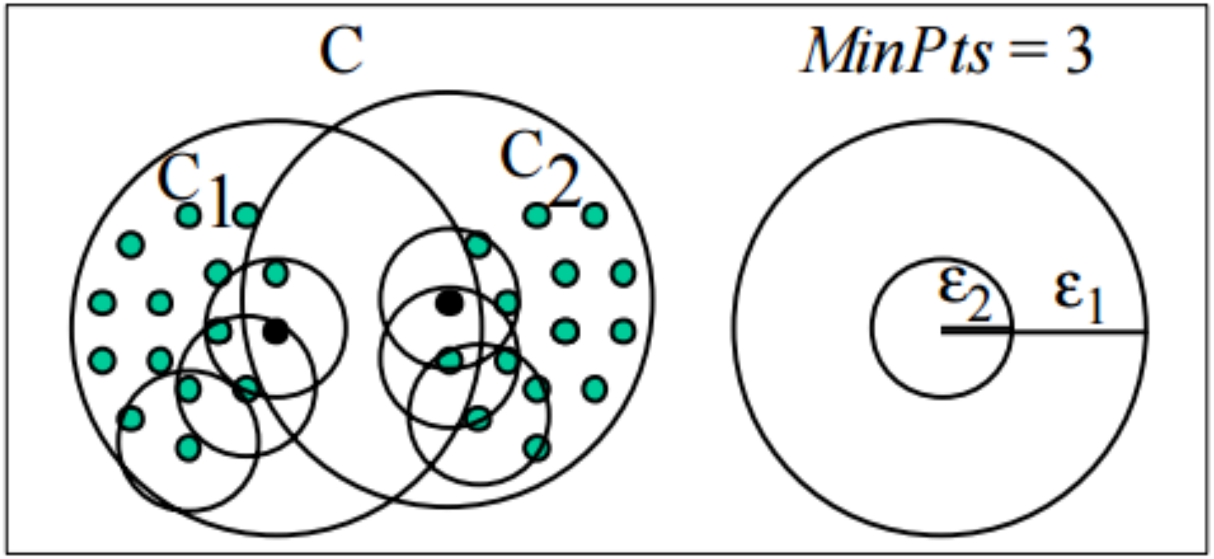
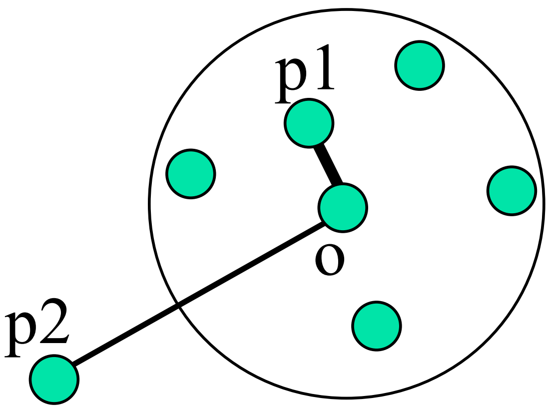


Figure 3. Illustration of "nested" density-based clusters

OPTICS: Some Extension from DBSCAN

- Core Distance (of o)
  - Distance to make the object a core
- Reachability Distance (of p from o)
  - r(p, o) = max(core-distance(o), d(o, p))
  - Ex: r(p1, o) = 2.8cm, r(p2,o) = 4cm



MinPts = 6
ε = 3 cm

OPTICS: A Cluster-Ordering Method (1999)

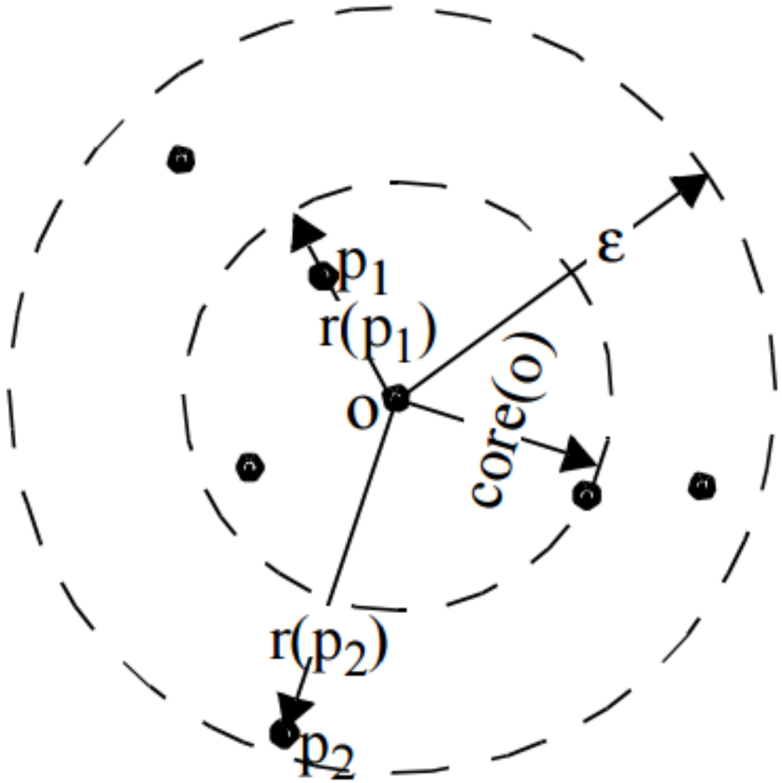
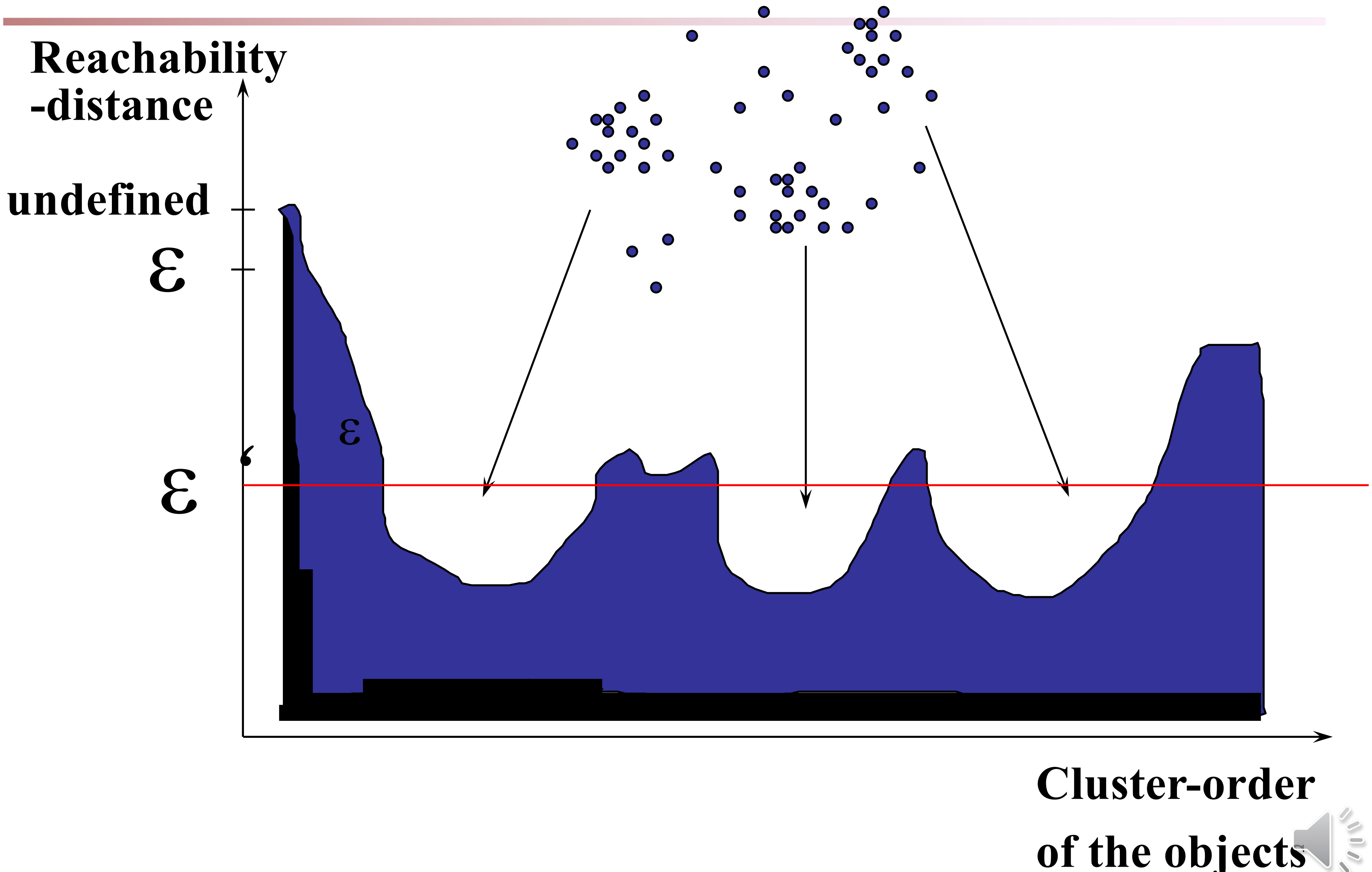
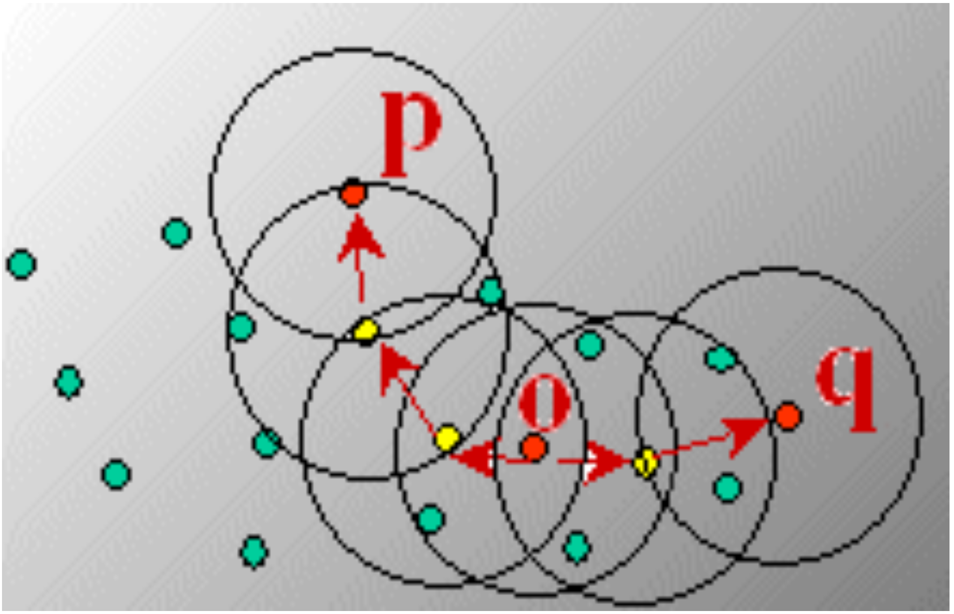
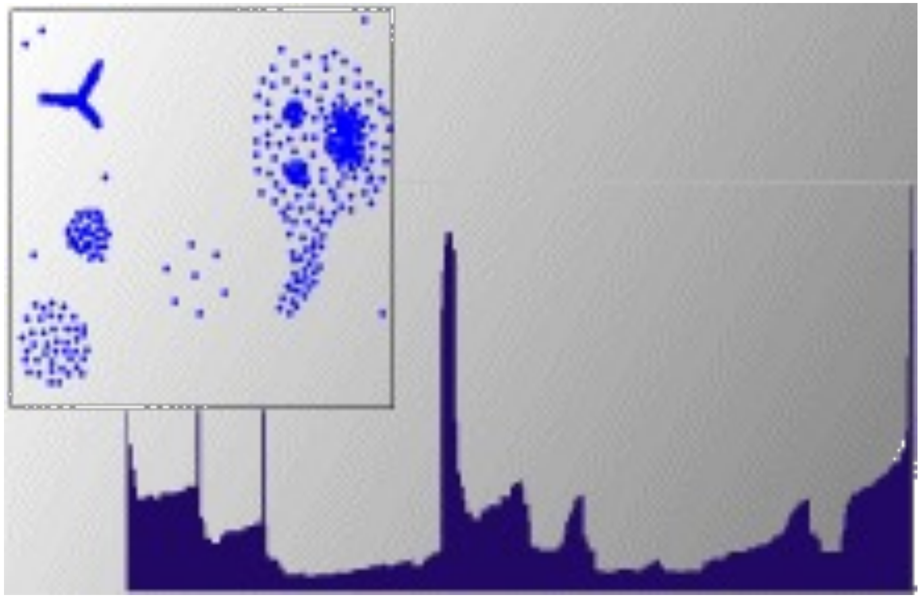


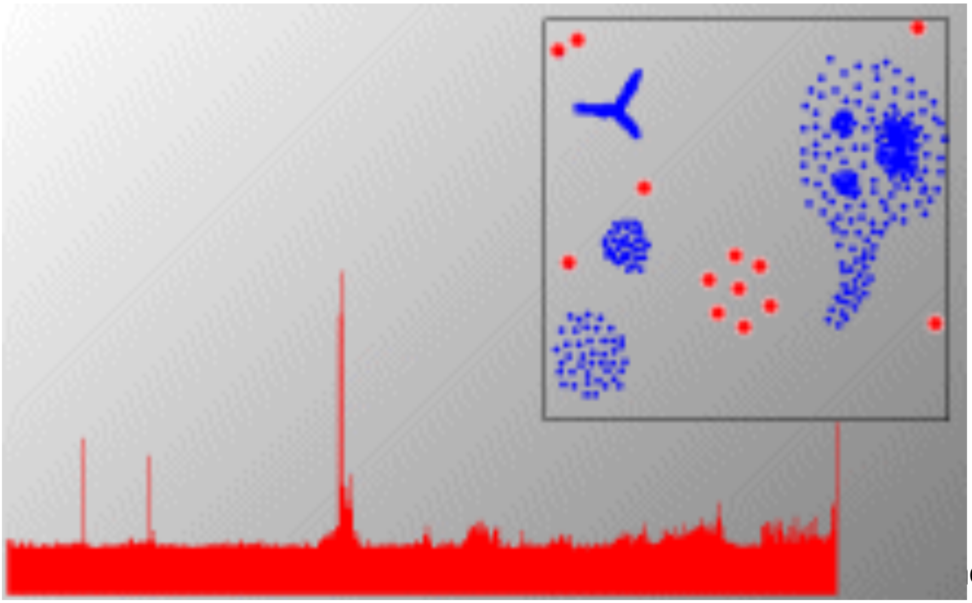
Figure 4. Core-distance(o), reachability-distances r(p_1,o), r(p_2,o) for *MinPts*=4

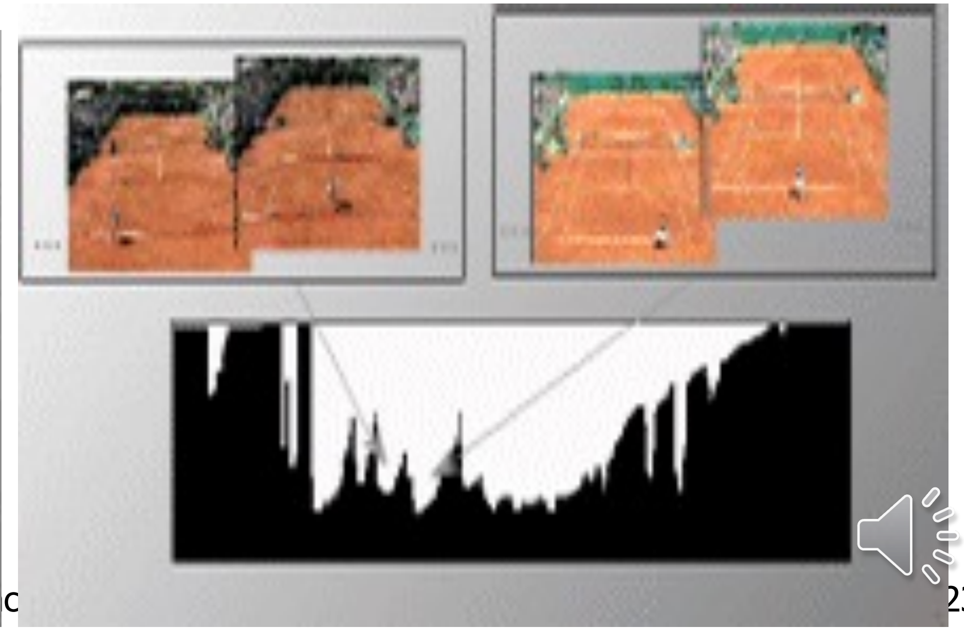

Reachability
-distance
undefined
ε
ε
ε'
Cluster-order
of the objects

Density-Based Clustering: OPTICS & Its Applications









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 objects in the database w.r.t. its density-based clustering structure
  - This cluster-ordering contains the information equivalent to different density-based clustering structure corresponding to a broad range of parameter settings (*Eps*)
  - Good for both automatic and interactive cluster analysis, including finding intrinsic clustering structure
  - Can be represented graphically or using visualization techniques

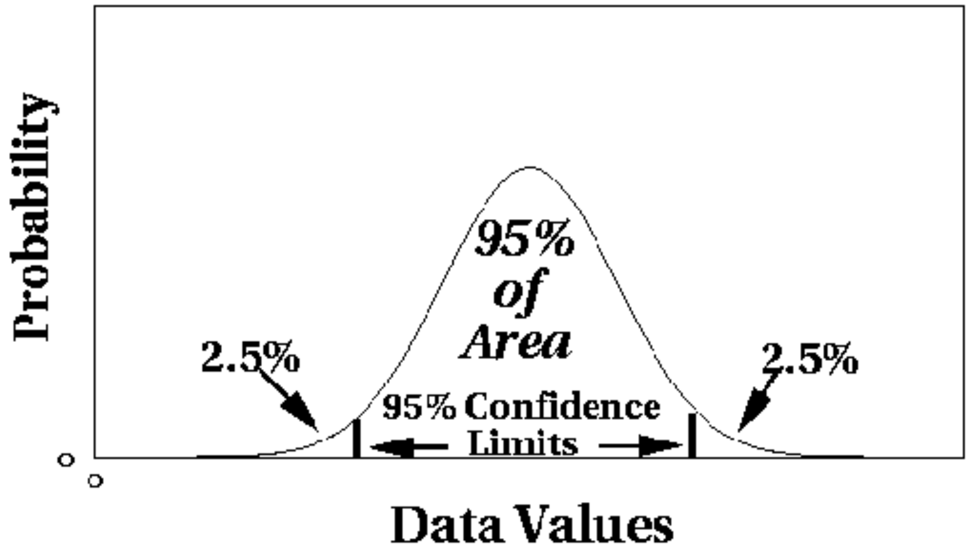
Chapter 7. Cluster Analysis

1. What is Cluster Analysis?
2. Types of Data in Cluster Analysis
3. A Categorization of Major Clustering Methods
4. Partitioning Methods
5. Hierarchical Methods
6. Density-Based Methods
7. Outlier Analysis
8. Summary

What Is Outlier Discovery?

- What are outliers?
  - The set of objects are considerably dissimilar from the remainder of the data
  - Example: Sports: Michael Jordon, Wayne Gretzky, ...
- Problem: Define and find outliers in large data sets
- Applications:
  - Credit card fraud detection
  - Telecom fraud detection
  - Customer segmentation
  - Medical analysis

Outlier Discovery: Statistical Approaches



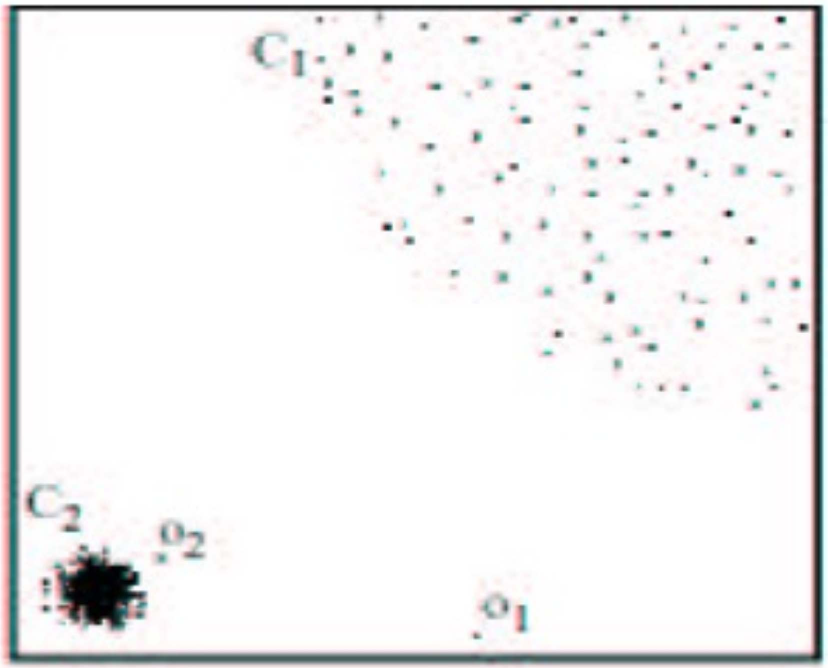
- Assume a model underlying distribution that generates data set (e.g. normal distribution)
- Use discordancy tests depending on
  - data distribution
  - distribution parameter (e.g., mean, variance)
  - number of expected outliers
- Drawbacks
  - most tests are for a *single attribute*
  - In many cases, data distribution may not be known

Outlier Discovery: Distance-Based Approach

- Introduced to counter the main limitations imposed by statistical methods
  - We need multi-dimensional analysis without knowing data distribution
- Distance-based outlier: A *DB(p, D)-outlier* is an object O in a dataset T such that at least a fraction ***p*** of the objects in T lies at a distance greater than ***D*** from O
- Algorithms for mining distance-based outliers
  - Index-based algorithm
  - Nested-loop algorithm
  - Cell-based algorithm

Density-Based Local Outlier Detection

- Distance-based outlier detection is based on global distance distribution
- It encounters difficulties to identify outliers *if data is not uniformly distributed*
- Ex. C_1 contains 400 loosely distributed points, C_2 has 100 tightly condensed points, 2 outlier points o_1, o_2
- Distance-based method cannot identify o_2 as an outlier
- Need the concept of *a local outlier*



- Local outlier factor (LOF)
  - Assume outlier is not crisp
  - Each point has a LOF

Chapter 7. Cluster Analysis

1. What is Cluster Analysis?
2. Types of Data in Cluster Analysis
3. A Categorization of Major Clustering Methods
4. Partitioning Methods
5. Hierarchical Methods
6. Density-Based Methods
7. Clustering High-Dimensional Data
8. 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
- Outlier detection and analysis are very useful for fraud detection, etc. and can be performed by statistical, distance-based or deviation-based approaches
- There are still lots of research issues on cluster analysis

Problems and Challenges

- Considerable progress has been made in scalable clustering methods
  - Partitioning: k-means, k-medoids, CLARANS
  - Hierarchical: BIRCH, ROCK, CHAMELEON
  - Density-based: DBSCAN, OPTICS, DenClue
  - Constraint-based: COD, constrained-clustering
- Current clustering techniques do not address all the requirements adequately, still an active area of research