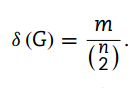
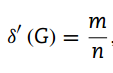
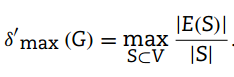
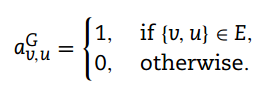
**Graph Cluster**

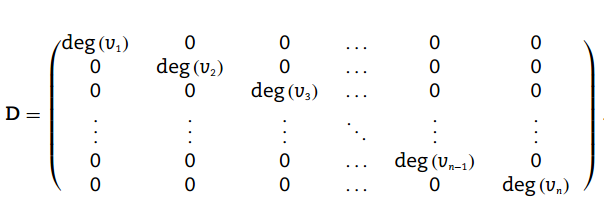
1. basic definitions
   1. Computational complexity
   2. Approximation algorithms
   3. Graph theory
      1. Density (For n ∈ {0, 1}, we set δ (G) = 0. A graph of density one is called complete)

* + 1. Neighbourhood Γ (v)
    2. Adjacency matrix AG



* + 1. Degree martrix D

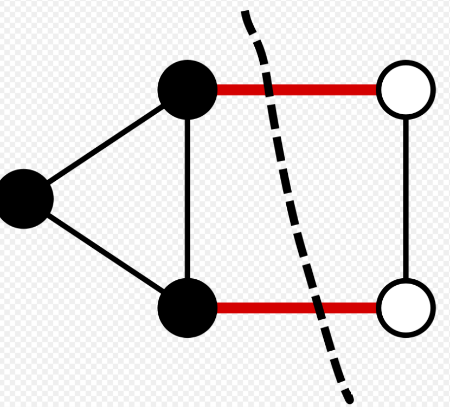
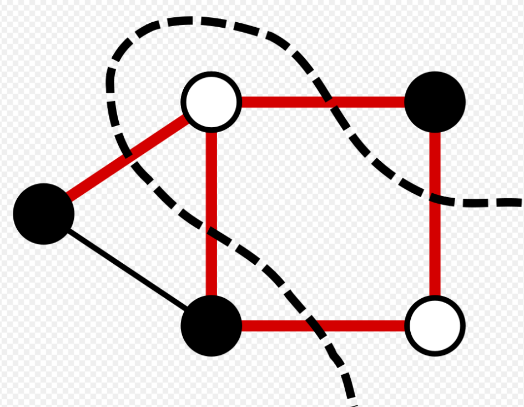


* + 1. cut size is the number of edges that connect vertices in S to vertices in V\S.

(c (S, V\S) = |{{v, u} ∈ E | u ∈ S, v ∈ V\S}|)

deg (S) = Σv∈Sdeg (v)

（以顶点为准，切分成两个不相交的子集）

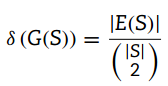
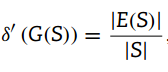
 

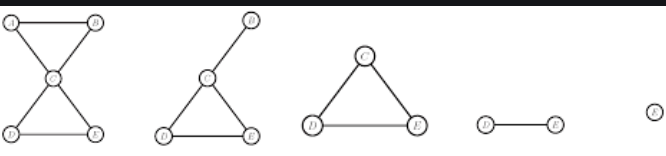
* + 1. If such a path exists, v and u are connected. The path is simple if no vertex is repeated.
    2. The length of a path is the number of edges on it, and the distance between v and u is the length of the shortest path connecting them in G.
    3. A graph is connected if there exist paths between all pairs of vertices.
    4. A cycle is a simple path that begins and ends at the same vertex. (A graph that contains no cycle is acyclic and is also called a forest. A connected forest is called a tree)
    5. A connected acyclic subgraph that includes all vertices is called a spanning tree of the graph. (If the edges are assigned weights, the spanning tree with smallest total weight is called the minimum spanning tree. (not only))
    6. An induced subgraph of a graph G = (V, E) is the graph with the vertex set S ⊆ V with an edge set E(S) that includes all such edges {v, u} in E with both of the vertices v and u included in the set S. (An induced subgraph that is a complete graph is called a clique.)

（团：团内任意两个结点有边（完全图、至少两个结点））

（诱导子图：首先是个子图，包含子图中的点图之间的所有边（不一定时完全图，可以一个结点））

（诱导子图是个多于一个结点的完全图 = 团）



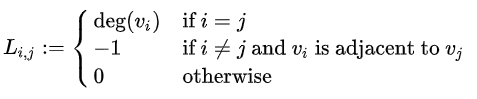
* + 1. Two graphs Gi = (Vi, Ei) and Gj = (Vj, Ej) are isomorphic if there exists a bijective (one-to-one) mapping f : Vi → Vj (called an isomorphism) such that {v, w} ∈ Ei if and only if {f(v), f(w)} ∈ Ej.

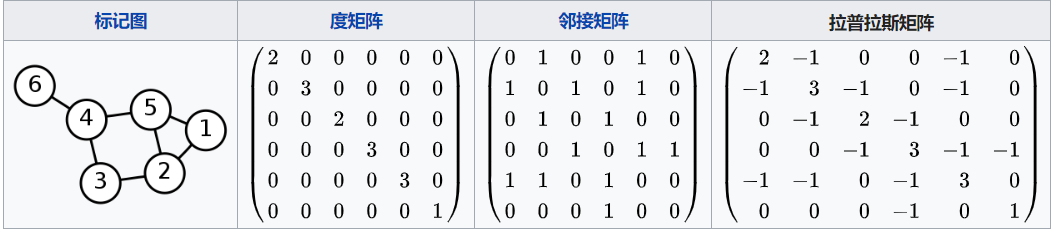
（对于映射顺序，如果两个图是同构的，则两个图的邻接矩阵是相同的，则谱相同 cospectral，即特征值相同）

（如果G1和G2同构，G1具有某性质，则G2也具有此性质）

（分类：精确图完全同构、精确子图同构、不精确图完全同构、不精确子图同构（后面三者是NP-Complete问题））

* + 1. The spectrum of a graph G = (V, E) is defined as the list of eigenvalues.
    2. Laplacian matrix L = D − AG

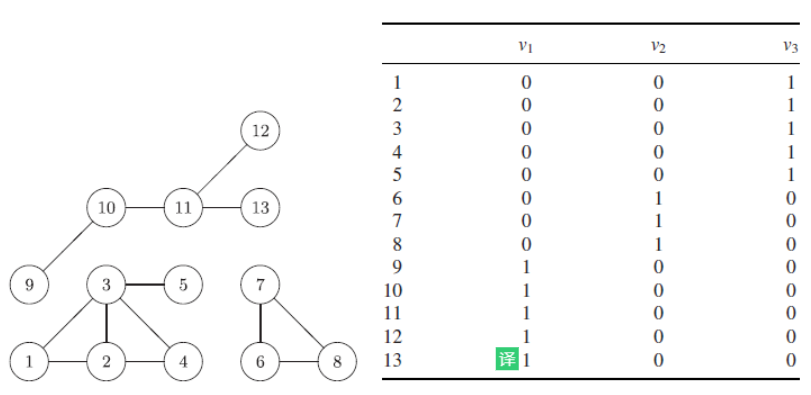




（谱聚类是利用相似矩阵或其他派生矩阵的结构特征，将样本划分到不相交类别中，并使类内样本相似度很高，而类别间样本相似度较低的一类技术，是一种启发式的聚类算法）

（拉普拉斯矩阵0特征值的个数为3个，这就表明其由3个相互分离的部分，特征向量矩阵的行表示出结点属于那一部分）

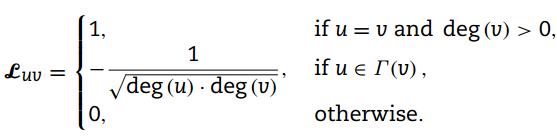
（最小特征值为零，表示该图包含一个连通分量）



* + 1. Normalized Laplacian

(eigenvalues∈[0，2], smallest eigenvalue is always zero, as the matrix is singular, and the corresponding eigenvector is simply a vector with each element being the square-root of the degree of the corresponding vertex.)





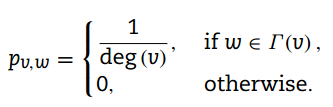
* + 1. Nonisomorphic graphs can share the same spectrum
    2. Graphs that have the same spectrum are called cospectral
    3. When the equality of the sets of pairwise distinct eigenvalues holds, but the multiplicities do not coincide, the graphs are weakly cospectral. (俩图特征值是相同的，但多重特征值不一样)
    4. minimum value of the Rayleigh quotient is the eigenvalue itself.
    5. The right eigenvector associated to the second-smallest eigenvalue of the Laplacian matrix is called a Fiedler vector, named after Fiedler

（使用 Fiedler 向量 w 将图分区为两个子图。如果一个节点在 w 中具有正值，则该节点将分配至子图 A。否则，该节点将分配至子图 B。这种做法称为符号切割或零阈值切割。符号切割最大限度减小了切割权重，该权重受限于图的任意非平凡切割的权重上界和下界。）

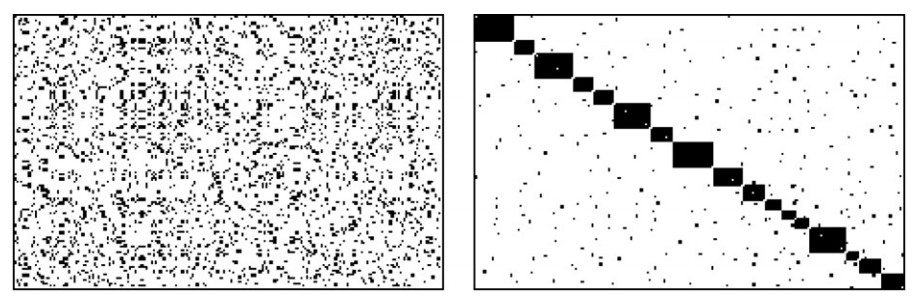
（Fiedler向量，即对应于图的拉普拉斯算子的第二小特征值的特征向量，在许多应用中起着至关重要的作用，包括矩阵重排序、图分区，蛋白质分析、数据挖掘、机器学习和网络搜索 ）

（图的最稀疏切割可以通过Cheeger不等式的拉普拉斯算子的第二个最小特征值来近似）

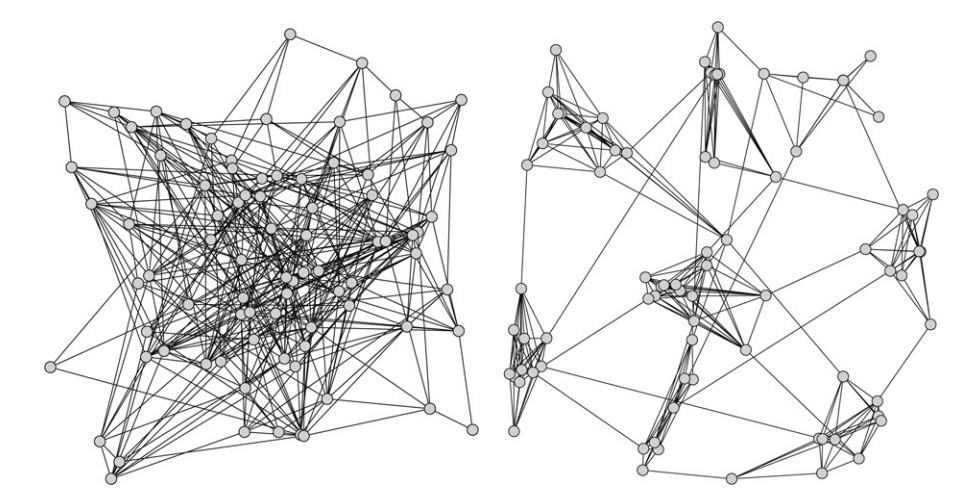
* 1. Markov chains
     1. A Markov chain is a stochastic process in which future states only depend on the current state, not the past. (No Memory)
     2. The probabilities for moving to another state from current state form the transition matrix of the Markov chain.
     3. A weighted directed graph where each state corresponds to a vertex, each edge corresponds to a transition that has nonzero probability and the edge weight is the probability in question.
     4. An unweighted graph, when one moves from one vertex to another choosing a neighbouring vertex uniformly at random, the transition matrix that results is the normalized adjacency matrix D−1AG of the graph G.

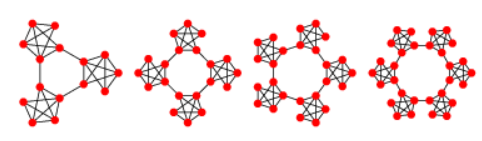


1. the different definitions of clusterings and clusters
   1. an adjacency matrix of a graph with n = 210 vertices and m =1505 edges



* 1. Such computations in turn enable efficient algorithms for graph partitioning, as the graph partitioning problem can be written in the form of a set of linear equations. (minimize the number of edges that cross from one subgroup of vertices to another, usually posing limits on the number of groups as well as to the relative size of the groups)
  2. uniform random graph (each of the n 2 possible edges is included in the graph with probability p, considering each pair of vertices independently degree distribution is Poissonian) and relaxed caveman structure (linking together a ring of small complete graphs called “caves”)

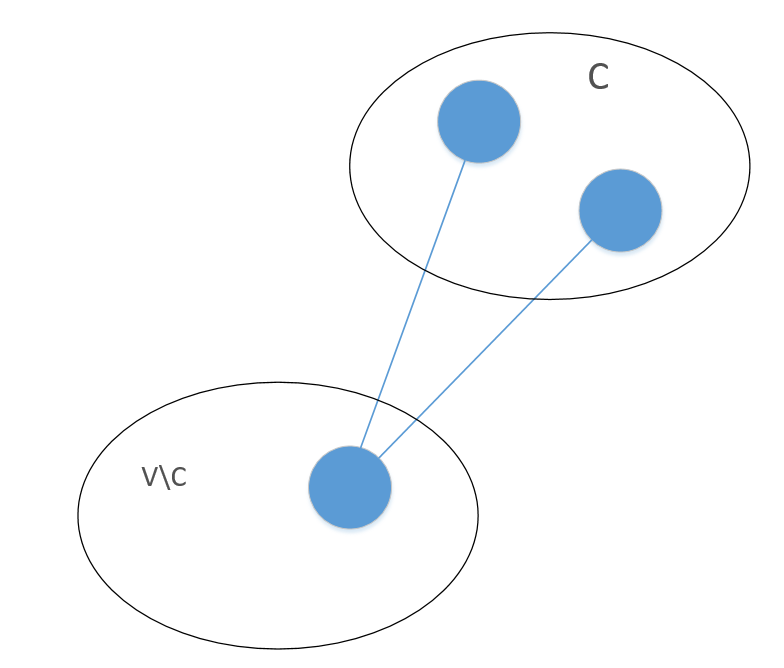




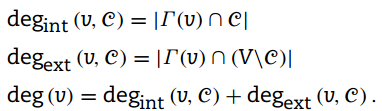
(artificial input data for evaluating and benchmarking clustering algorithms.)

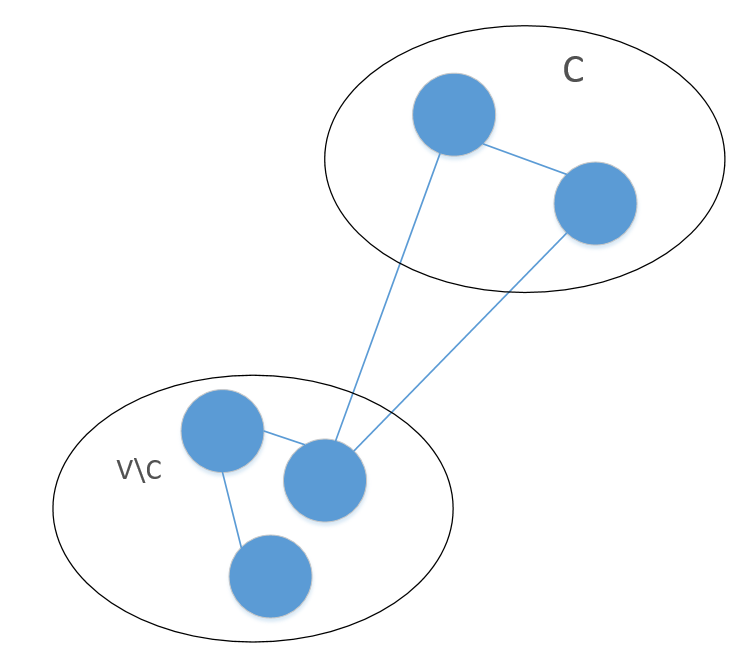
(the relaxed caveman model incorporates the generation of a hierarchy and the possibility to generate unequal-sized clusters)

* 1. The planted l-partition model: a graph is generated with n = l · k vertices that are partitioned into l groups each with k vertices. (each pair of vertices that are in the same group share an edge with the higher probability p, whereas each pair of vertices in different groups shares an edge with the lower probability r (r<p))
  2. cluster properties
     1. each cluster should intuitively be connected: there should be at least one.
     2. If a vertex u cannot be reached from a vertex v, they should not be grouped in the same cluster. Two vertices v and u in C also need to be connected by a path that only visits vertices included in C.

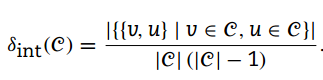


* + 1. when clustering a disconnected graph with known components, the clustering should usually be conducted on each component separately, unless some global restriction on the resulting clusters is imposed.
    2. We classify the edges incident on v ∈ C into two groups: internal edges that connect v to other vertices also in C, and external edges that connect v to vertices that are not included in the cluster C. (degext (v) = 0 implies that C containing v could be a good cluster)

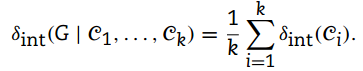




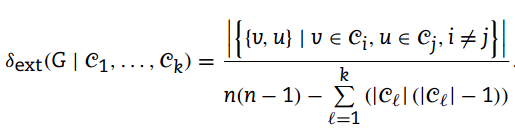
* + 1. the internal or intra-cluster density



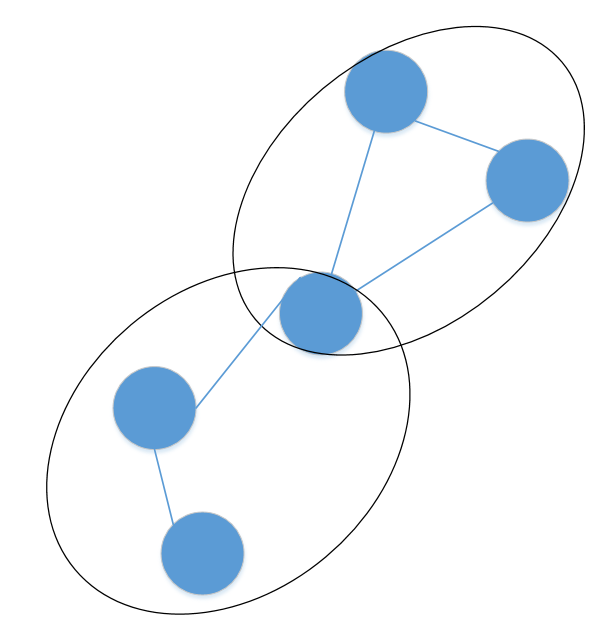
* + 1. The intercluster density of a given clustering of a graph G



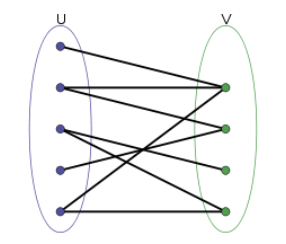
* + 1. The external or inter-cluster density normalized to the range [0, 1]



* + 1. the internal density of a good clustering should be notably higher than the density of the graph δ (G) and the intercluster density of the clustering should be lower than the graph density
    2. the loosest possible definition of a graph cluster is that of a connected component, and the strictest definition is that each cluster should be a maximal clique
    3. It is not always clear whether each vertex should be assigned fully to a cluster or could it instead have different “levels of membership” in several clusters?



* + 1. fuzzy clustering algorithms (In fuzzy clustering, the data points can belong to more than one cluster) (A fuzzy graph GR = (V, R) is composed of a set of vertices and a fuzzy edge-relation R that is reflexive and symmetrical together with a membership function µR assigns to each fuzzy edge a level of “presence” in the graph) （在计算上，创建模糊边界要比在一个集群上建立一个点要容易得多）
  1. Representations of clusters
     1. Bipartite graphs



(grouping the customers by the types of products they purchase or grouping products purchased by the same people)

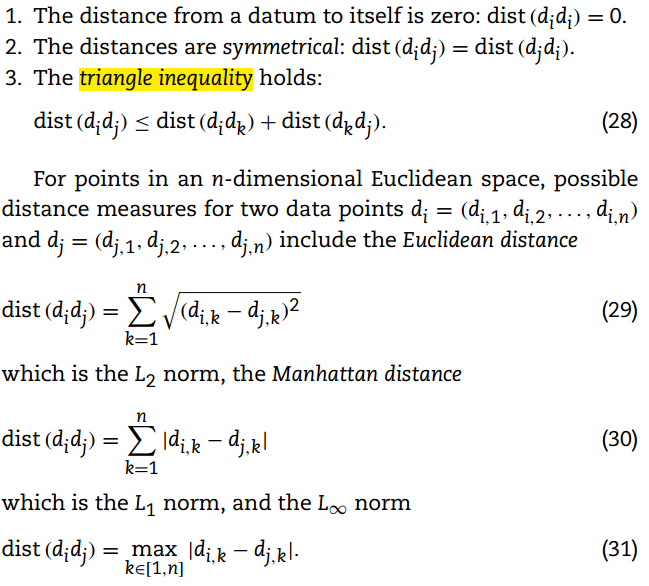
(the overlap of the neighbourhoods the one side of the graph reflects the similarity of the vertices of the other side and vice versa)

* + 1. Directed graphs

(Web graphs are directed graphs formed by web pages as vertices and hyperlinks as edges)

(Clustering of web pages can help identify topics and group similar pages)

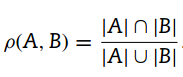
1. Measures for identifying clusters（4）
   1. How to identify a good cluster?
      1. compute some values for the vertices and then classify the vertices into clusters based on the values obtained
      2. compute a fitness measure over the set of possible clusters and then choose among the set of cluster candidates those that optimize the measure used
   2. Vertex similarity (not only well connected but also similar to each other)
      1. Distance and similarity measures (对比结点的属性，例如书本的作者、内容等)



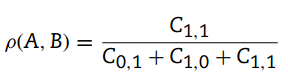
(A typical example of a nonEuclidean space is that formed by vector representations of textual data)

* + 1. the Jaccard index (cluster validity index)

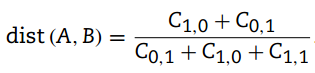
(对比A,B set之间的相似度)



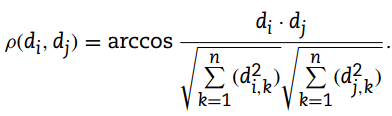
Jaccard similarity coefficient



Jaccard distance = 1-ρ(A,B)

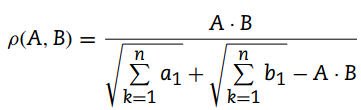


* + 1. cosine similarity (angle in [0, π)) (vectors di,dj，假如类似，则角度越小)



* + 1. the Tanimoto coefficient (cos similarity + jaccard coefficient)

(The cosine similarity was extended to the coincide with the Jaccard similarity for n-dimensional binary vectors)



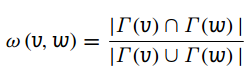
* + 1. Adjacency-based measures （对比结点外在属性，例如书本拥有的用户是否一样）

(the vertices lack additional properties make the computation of a similarity matrix)

(using adjacency matrix or computation to drive similarity measures)

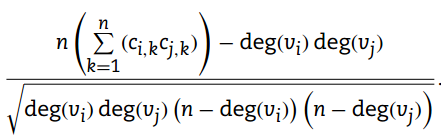
(vertex-similarity measures based on the structural properties of the graph instead of some application-specific properties imposed on the vertices)

* + 1. the overlap of their neighbourhoods ([0, 1])



* + 1. Pearson correlation (of col or row)(衡量两个数据集合是否在一条线上面，它用来衡量定距变量间的线性关系)(cosine similarity的扩展)

(a modified adjacency matrix C = AG + I)



(This value can then be used as an edge weight ω(vi, vj) to construct a symmetrical similarity matrix)

(It reaches the value 1 if and only if the two vertices have the same neighbourhood, and for neighbourhoods with no overlap the value is a negative number no less than −1, depending on the degrees of the vertices.)

* + 1. Connectivity measures

(be highly connected to each other in the same cluster)

(a good cluster, if they are at least connected by a short path, it is not absolutely necessary that two included vertices v and u are connected by a direct edge)

(a matrix containing the distance for each vertex pair could serve as a similarity matrix)

* + 1. threshold the path length

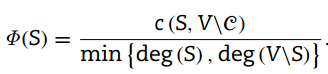
(all vertices in a cluster must be at distance at most k from

each other)

(set the threshold k by the diameter of the input graph which is the maximum distance over all pairs of nodes)

* 1. Cluster fitness measures
     1. the identification of clusters (by optimizing one or more cluster fitness measures)
     2. choosing between alternative clusterings (by comparing the values of one or more cluster fitness measures)
     3. comparing different clustering algorithms (by studying the values that the clusterings computed give for different cluster fitness measures)
     4. Density measures
     5. maximal subgraphs that have a density higher than a preset threshold (ξ ∈ [0, 1]) (|S|=k)
     6. p-quasi complete subgraph for p∈[0,1], if for all v∈V, deg(v) >= p(n-1)
     7. Cut-based measures (edge present, connectivity)
     8. Conductance (衡量cluster内容的重要指标)

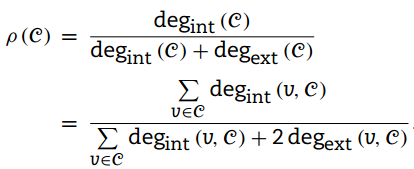
(normalized cut, cur ratio, sparsest cut problem)



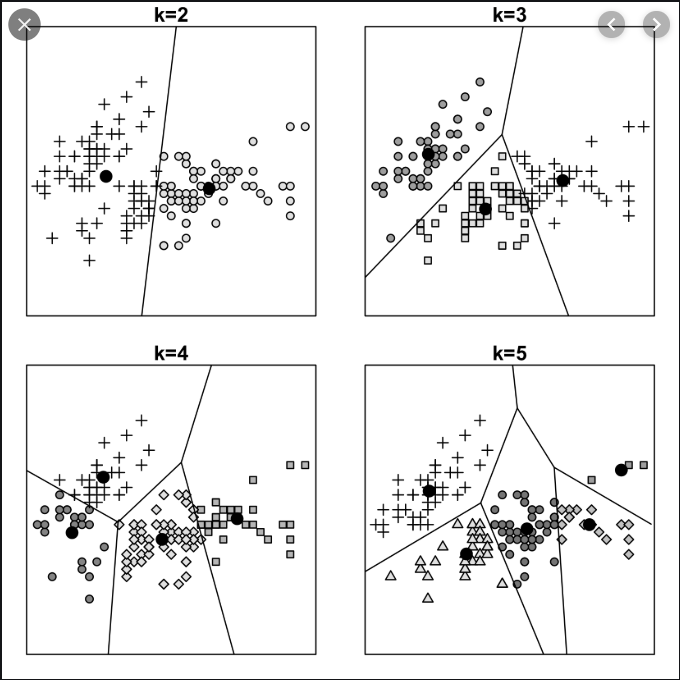
(c(S,V\C), cut size)

(the external degree is in fact the size of the cut (C, V\C).)

* + 1. Relative density (the ratio of the internal degree to the number of incident edges)



1. **global clustering（5）**
   1. Complexity
      1. the maximum intercluster distance is minimized
      2. A related problem is the minimum k-centre problem
      3. A weighted version of the k-centre problem
      4. minimum k-median problem
      5. k-means algorithm (NP-hard)

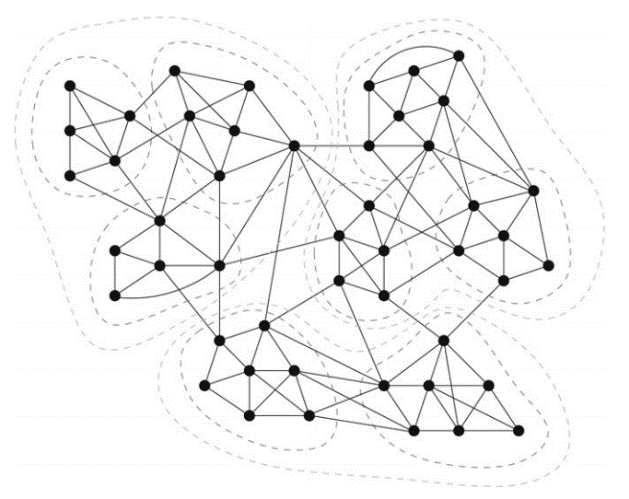


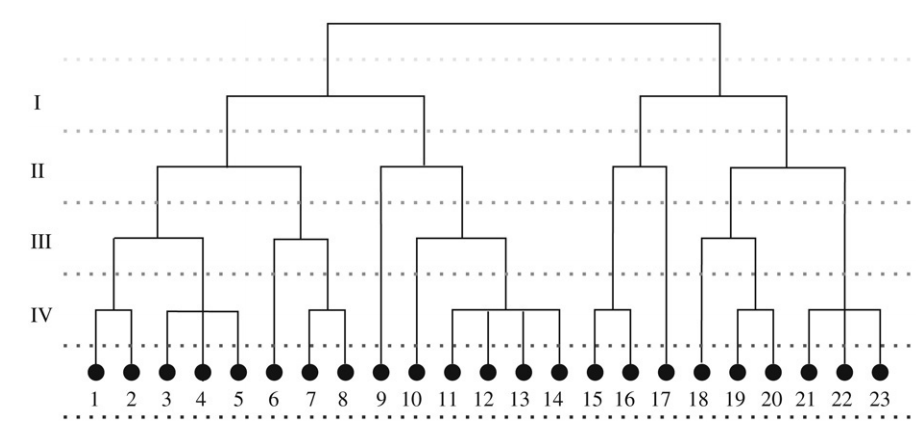
* 1. Iterative or online computation
     1. Time, Not scale
     2. Operates one datum at a time -> operate online
     3. Time, dynamically, relying on some threshold value, merging it to an existing cluster
     4. incremental clustering (incrementally incorporating newly added data into the existing clustering) (web page classification)
     5. the nearest neighbour of the point (the distance measure used

should preferably incorporate some structural information on

connectivity among the vertices）

* + 1. the order in which the data are processed may affect the resulting clusters (immutable)
    2. memory consumption
  1. Hierarchical clustering (k ≥ 3 the problem is NP-hard)





* 1. top-down or divisive
     1. Cuts
     2. Maximum flow
     3. Spectral methods
     4. Betweenness
     5. Voltage and potential
     6. Markov chains and random walks
  2. bottom-up or agglomerative

(a similarity measure is used to the vertices to be merged into a cluster)

(iterative merging typically continues until some threshold or a desired number of clusters is reached.)

(initial clusters using information on vertex degrees and then agglomeratively combines the initial clusters until an agreeable clustering is achieved.)

* + 1. the pairwise nearest neighbours method
    2. the merging criterion
    3. modularity-optimizing approach (using spectral properties to construct the full cluster hierarchy and then select a clustering from the resulting tree maximizing modularity.)

1. **local clustering（6）**

(For large enough graphs, even sparsity does not help much)

(clusters can be computed one at a time based on only partial views of the graph topology.)

(parallel computation)

(applying some majority-vote rule or a quality measure to combine the local clusters into a global clustering) (fuzzy global clustering)

(seed vertices)

* 1. Definition of locality in graphs

(define what information about an input graph we consider to be locally available.)

(allow a local algorithm to remember any adjacency information that has already been seen,)

* 1. Local search

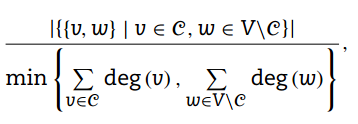
(to find near-optimal solutions in limited region)

(Each solution candidate is represented by a state)

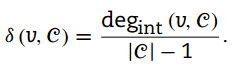
(define a fitness function repeatedly evaluated for different solution candidates.)

(the search algorithm always remembers at least the best state visited and the associated fitness.)

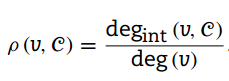
* + 1. Hill-climbing
    2. deterministic and probabilistic versions of tabu search
    3. Simulated annealing
  1. Fitness functions
     1. use local search to optimize the a weighted version of the Cheeger ratio



* + 1. This measure indicates how densely v is connected to C and it should give a high value if C is a good cluster for v

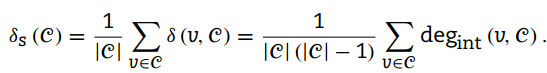


* + 1. make sure that the vertex is not densely connected to other parts of the graph

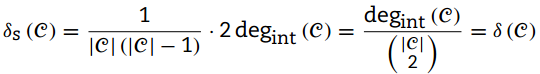


(If both of the above measures have a high value, we can assume v to be correctly classified into C)

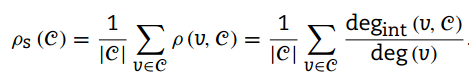
* + 1. The quality of a given cluster



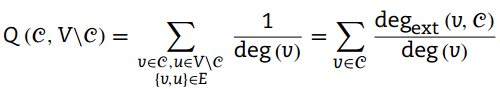
* + 1. obtaining exactly the local density of the subgraph induced by the vertex set C



* + 1. the local density can also be interpreted as the probability that two included vertices are connected
    2. the introversion of a cluster C (内向性)



* + 1. calculate the capacity of the cut (C, V\C) for a blind



1. random walkcomparing, evaluating and benchmarking