Heidelberg University
Institute of Computer Science
Database Systems Research Group

Lecture: Complex Network Analysis

Prof. Dr. Michael Gertz

Assignment 8
Clustering and Modularity

`https://github.com/nilskre/CNA_assignments`

Team Member: Patrick Günther, 3660886,
Applied Computer Science
rh269@stud.uni-heidelberg.de

Team Member: Felix Hausberger, 3661293,
Applied Computer Science
eb260@stud.uni-heidelberg.de

Team Member: Nils Krehl, 3664130,
Applied Computer Science
pu268@stud.uni-heidelberg.de

Problem 8-2 Hierarchical Clustering

For hierarchical clustering, we need to define the similarity of two disjoint clusters $X = \{x_1,\dots x_i\} \subset V$ and $Y = \{y_1,\dots,y_j\} \subset V$, based on the similarity of two nodes $s(v_i,v_j)$ with $v_i,v_j \in V$.

Popular linkage strategies for agglomerative hierarchical clustering include

	Similarity $s(X,Y)$	Distance $d(X,Y)$
Single-linkage clustering	$\max_{x\in X,y\in Y} s(x,y)$	$\min_{x\in X,y\in Y} d(x,y)$
Complete-linkage clustering	$\min_{x\in X,y\in Y} s(x,y)$	$\max_{x\in X,y\in Y} d(x,y)$

Given disjoint clusters X, X', and Y, we can alternatively compute the similarity $s(X\cup X',Y)$ using

	Similarity $s(X\cup X',Y)$	Distance $d(X\cup X',Y)$
Single-linkage clustering	$\max\{s(X,Y),s(X',Y)\}$	$\min\{d(X,Y),d(X',Y)\}$
Complete-linkage clustering	$\min\{s(X,Y),s(X',Y)\}$	$\max\{d(X,Y),d(X',Y)\}$

1. Discuss why Single-linkage and Complete-linkage swap their definitions (i.e. *min* and *max*) when used with a similarity rather than a distance function.

   After the merge process of nodes into communities, there are three approaches for calculating the new similarities: single-linkage (take the maximum), complete-linkage (take the minimum) and average-linkage. If the distance between two nodes is small (min), the similarity is high (max). And vice versa if the distance between two nodes is high (max), the similarity is low (min). Complete-linkage uses the maximum distance (which is the minimum similarity). Single-linkage uses the maximum similarity (which is the minimum distance).

2. Given the following similarity matrix x^o_{ij} , based on the topological overlap (cf. slide 9-22), perform agglomerative hierarchical clustering, using single-linkage.

$$x^o_{ij} = \begin{array}{c} \begin{matrix} A & B & C & D & E & F \end{matrix} \\ \begin{pmatrix} 0 & 1 & 1/2 & 1 & 1 & 0 \\ 1 & 0 & 1/3 & 1 & 1 & 0 \\ 1/2 & 1/3 & 0 & 1/3 & 1/3 & 1 \\ 1 & 1 & 1/3 & 0 & 1 & 0 \\ 1 & 1 & 1/3 & 1 & 0 & 0 \\ 1/2 & 0 & 1 & 0 & 0 & 0 \end{pmatrix} \end{array} \begin{matrix} A \\ B \\ C \\ D \\ E \\ F \end{matrix} \tag{1}$$

   Recall that for hierarchical clustering, you need to:

   - Find the maximum similarity $s(i,j)$,

- Merge the corresponding clusters,
- Update the similarity matrix by merging rows/columns i and j according to the linking strategy.

Begin by assigning each node to a separate cluster (each node is its own cluster), and repeat the above procedure until only one cluster remains. Write out the similarity matrix after each iteration and draw the resulting dendrogram. During each iteration, multiple sets may have the same maximum similarity. In this case, you may still only merge two clusters at a time. In this case, also prefer the merge of two smaller clusters over a bigger one, i.e., prefer to merge two clusters with cardinality 1 over the merge of clusters with size 2 and size 1, respectively.

---

single-linking strategy: take the minimum value

1. Merge A and B (maximum similarity 1)

$$x_{ij}^o = \begin{array}{c} \begin{array}{ccccc} A,B & C & D & E & F \end{array} \\ \left(\begin{array}{ccccc} 0 & 1/3 & 1 & 1 & 0 \\ 1/3 & 0 & 1/3 & 1/3 & 1 \\ 1 & 1/3 & 0 & 1 & 0 \\ 1 & 1/3 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{array}\right) \end{array} \begin{array}{c} A,B \\ C \\ D \\ E \\ F \end{array} \tag{2}$$

Cluster: (A,B),C,D,E,F

2. Merge C and F (maximum similarity 1)

$$x_{ij}^o = \begin{array}{c} \begin{array}{cccc} A,B & C,F & D & E \end{array} \\ \left(\begin{array}{cccc} 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{array}\right) \end{array} \begin{array}{c} A,B \\ C,F \\ D \\ E \end{array} \tag{3}$$

Cluster: (A,B),(C,F),D,E

3. Merge D and E (maximum similarity 1)

$$x_{ij}^o = \begin{array}{c} \begin{array}{ccc} A,B & C,F & D,E \end{array} \\ \left(\begin{array}{ccc} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{array}\right) \end{array} \begin{array}{c} A,B \\ C,F \\ D,E \end{array} \tag{4}$$

Cluster: (A,B),(C,F),(D,E)

4. Merge (A,B) and (D,E) (maximum similarity 1)

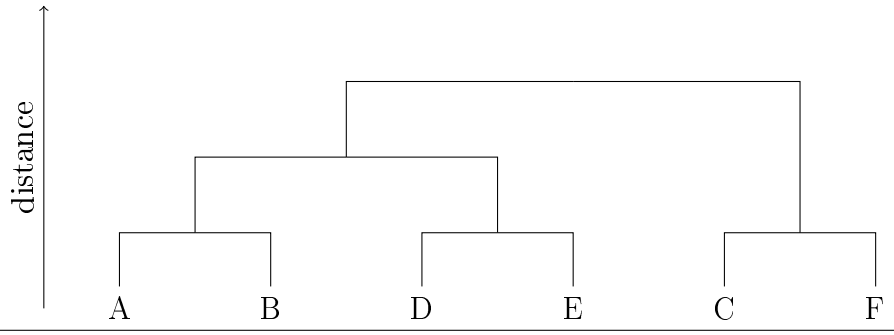
$$x_{ij}^{o} = \begin{array}{c} \begin{matrix} A,B,D,E & C,F \end{matrix} \\ \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} \end{array} \begin{matrix} A,B,D,E \\ C,F \end{matrix} \tag{5}$$

Cluster: (A,B,D,E),(C,F)

5. Merge (A,B,D,E) and (C,F) (maximum similarity 1/3)

results in one cluster containing all nodes Cluster: (A,B,D,E,C,F)

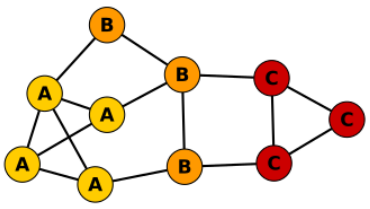
Dendrogram:



Problem 8-3 Modularity

Consider a simple, undirected graph G with L links between n nodes that are partitioned into n_c disjoint communities C. Let L_c be the number of links inside a given community c. Furthermore, let k_c be the sum of all degrees of nodes in community c (including edges that leave the community). Then the modularity M is defined as

$$M(G,C) := \sum_{c=1}^{n_c} (\frac{L_c}{L} - (\frac{k_c}{2L})^2) \tag{6}$$



1. Compute the modularity of the graph.

   - $L = 15$
   - $n = 10$
   - $n_c = 3$
   - $L_A = 5; L_B = 2; L_C = 3$
   - $k_A = 3 + 3 + 3 + 4 = 13; k_B = 2 + 4 + 3 = 9; k_C = 3 + 2 + 3 = 8$

   Plug the values into the given formula:

   $$M(G, C) := (\frac{5}{15} - (\frac{13}{2 * 15})^2) + (\frac{2}{15} - (\frac{9}{2 * 15})^2) + (\frac{3}{15} - (\frac{8}{2 * 15})^2) = 0.318 \quad (7)$$

2. Give proof that for every partitioning of every simple graph, it always holds that $M \leq 1$.

   For each component, its links within itself can be at most $L_c = L$. In this case, the component would hold all the links of the graph. For this component $\frac{L_c}{L}$ would be 1. Since there is the term ${\frac{k_c}{2L}}^2$ subtracted from that, the summand for this component in the equation cannot be greater than 1. All other components remaining in the graph cannot have any links, which means that these are just unconnected single nodes. This means that their summand in the equation (ignoring the division by zero) would be 0. The whole equation thus cannot be greater than 1.

   For other cases where there are multiple components containing links, the total number of L_c for all components is still limited by L. Thus it has to hold that $\sum_{c=1}^{n_c} (\frac{L_c}{L}) \leq 1$. Since the other term in each summand $({\frac{k_c}{2L}}^2)$ can only decrease the overall sum, $M \leq 1$ has to hold.

   For the trivial case of all nodes being disconnected, $M = 0$ (ignoring the division by 0).