**Ludwig-Maximilians-Universität München**
**Institut für Informatik**
Prof. Dr. Thomas Seidl
Max Berrendorf, Julian Busch

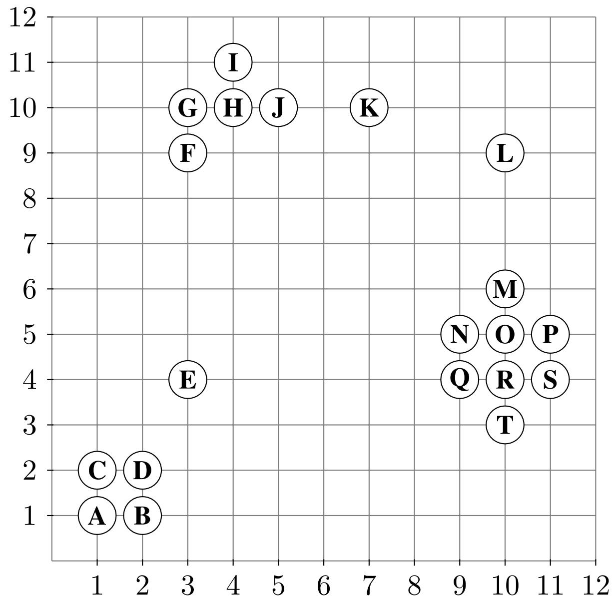
# Knowledge Discovery and Data Mining I

WS 2018/19

## Exercise 7: Agglomerative Clustering, OPTICS, Clustering Evaluation

### Exercise 7-1 Hierarchical Clustering

Given the following data set:




As distance function, use Manhattan Distance:

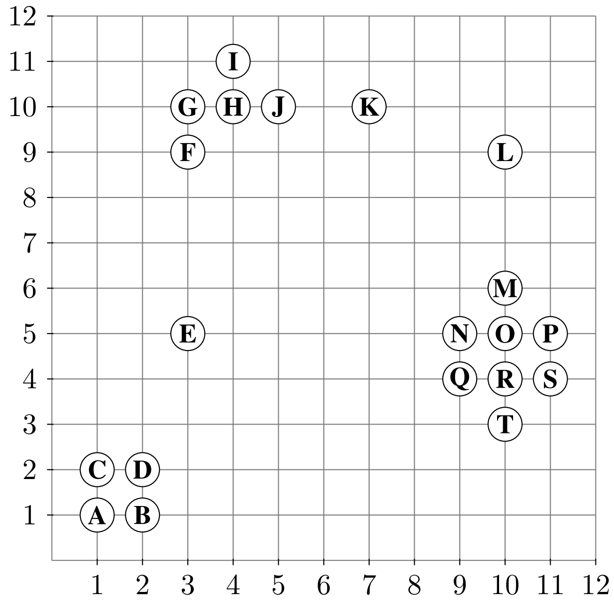
$$L_1(x, y) = |x_1 - y_1| + |x_2 - y_2|$$

Compute two dendrograms for this data set. To compute the distance of sets of objects, use

- the single-link method
- the complete-link method

**Hint**: With discrete distance values, nodes may have more than two children.

### Exercise 7-2 OPTICS




As distance function, use Manhattan distance $L_1(a,b) := |a_1 - b_1| + |a_2 - b_2|$.

Construct an OPTICS reachability plot for each of the following parameter settings. In case of a tie always proceed with the first candidate in alphabetical order.

(a) $\varepsilon = 5$ and *minPts* $= 2$

(b) $\varepsilon = 5$ and *minPts* $= 4$

(c) $\varepsilon = 2$ and *minPts* $= 4$

(d) $\varepsilon = \infty$ and *minPts* $= 4$

### Exercise 7-3 Efficient Evaluation of Clusterings

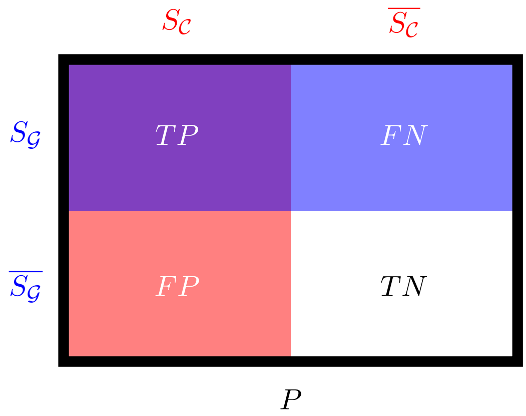
Let $D$ be a database of size $n := |D|$, and let $\mathcal{C}, \mathcal{G}$ be two partitionings of $D$. Furthermore, let $k := |\mathcal{C}|$ and $l := |\mathcal{G}|$ be the number of partitions, and assume that the contingency table is provided as a $(k \times l)$ matrix, where $N_{ij} = |C_i \cap G_j|$ denotes one cell in this table.

As in the lecture slides, let $P := \{(o,p) \in D^2 \mid o \neq p\}$ denote the set of all pairs, and $S_{\mathcal{C}} = \{(o,p) \in P \mid \exists C_i \in \mathcal{C} : \{o,p\} \subseteq C_i\}$ be the set of pairs that are contained in a common cluster $C_i$ in $\mathcal{C}$. In addition, $\overline{S_{\mathcal{C}}}$ denotes the complement of $S_{\mathcal{C}}$ in $P$, i.e. $\overline{S_{\mathcal{C}}} = P \setminus S_{\mathcal{C}}$. $S_{\mathcal{G}}$ and $\overline{S_{\mathcal{G}}}$ are used analogously.

Using these four sets, we can now define the

- *True Positives* (TP): Same labelling in $\mathcal{C}$ and same labelling in $\mathcal{G}$, i.e. $TP = |S_{\mathcal{C}} \cap S_{\mathcal{G}}|$
- *False Positives* (FP): Same labelling in $\mathcal{C}$, but different labelling in $\mathcal{G}$, i.e. $FP = |S_{\mathcal{C}} \cap \overline{S_{\mathcal{G}}}|$
- *False Positives* (FN): Different labelling in $\mathcal{C}$, but same labelling in $\mathcal{G}$, i.e. $FN = |\overline{S_{\mathcal{C}}} \cap S_{\mathcal{G}}|$
- *True Negatives* (TN): Different labelling in $\mathcal{C}$, and different labelling in $\mathcal{G}$, i.e. $TN = |\overline{S_{\mathcal{C}}} \cap \overline{S_{\mathcal{G}}}|$

The relation of these four sets and $S_{\mathcal{C}}$ as well as $S_{\mathcal{G}}$ is also visualised in the following Venn diagram:




For each of these cardinalities, provide a method to obtain the numbers solely from the contingency table, i.e. without explicitly enumerating set of all pairs (which requires $\mathcal{O}(n^2)$ time).

(a) $TP = |S_{\mathcal{C}} \cap S_{\mathcal{G}}|$,

(b) $FP = |S_{\mathcal{C}} \cap \overline{S_{\mathcal{G}}}|$,

(c) $FN = |\overline{S_{\mathcal{C}}} \cap S_{\mathcal{G}}|$,

(d) $TN = |\overline{S_{\mathcal{C}}} \cap \overline{S_{\mathcal{G}}}|$.

### Exercise 7-4 Mutual Information

Given are two clusterings of $D = \{A, \ldots, Z\}$:

- $\mathcal{C} = \{\{A, D, E, I, K, L, M, N, T\}, \{B, F, O, Q, R, S, X, Y\}, \{C, G, H, P, V\}, \{J, U, W, Z\}\}$.
- $\mathcal{G} = \{\{A, F, M, Y\}, \{B, E, H, N, O, Q, R, Z\}, \{C, G, U, W\}, \{D, I, K, L, P, T, V\}, \{J, S, X\}\}$.

(a) Setup the contingency table, i.e. compute the sizes $|C_i \cap G_j|$ for $i = 1, \ldots, 4$, and $j = 1, \ldots, 5$.

(b) Using the contingency table from (a), compute the entropy of $\mathcal{C}$ and $\mathcal{G}$, i.e. $H(\mathcal{C})$ and $H(\mathcal{G})$.

(c) Using the contingency table from (a), compute the mutual entropy $H(\mathcal{C} \mid \mathcal{G})$.

(d) Combine the results from (b) and (c) to obtain the normalised mutual information. What does this value tell about the two clusterings?