

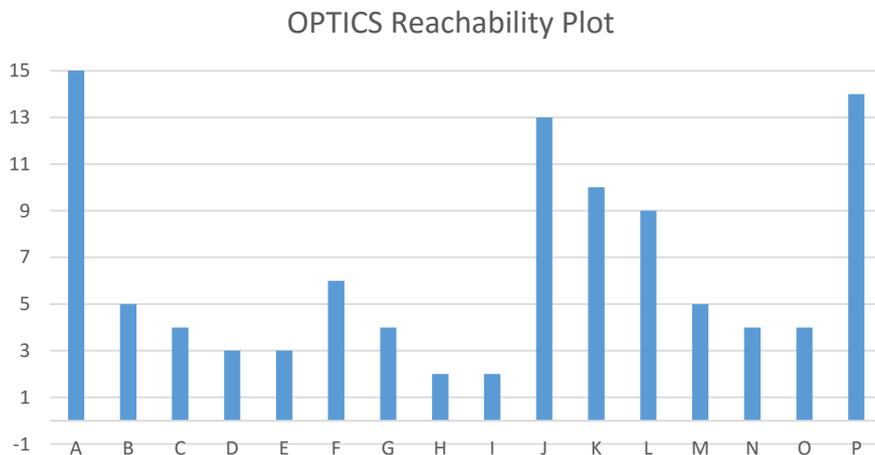
Knowledge Discovery in Databases  
WS 2017/18

Übungsblatt 9: Clustering und Outlier Detection

Besprechung: 18. und 19.01.2018

Aufgabe 9-1 OPTICS

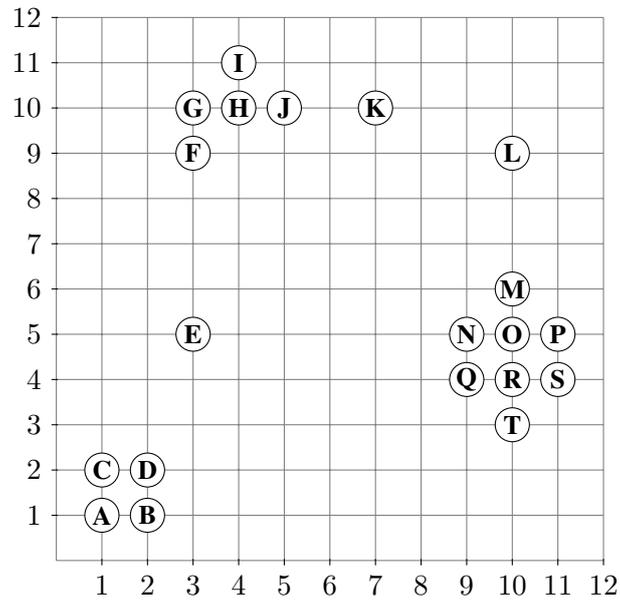
Gegeben ist der folgende OPTICS Erreichbarkeitsgraph.



- (a) Welches Clustering erhält man für  $\epsilon = 11$ ?
- (b) Welche Objekte sind Outlier bei  $\epsilon = 11$ ?
- (c) Welches Clustering erhält man für  $\epsilon = 4$ ?
- (d) Welche Objekte sind Outlier bei  $\epsilon = 4$ ?
- (e) Für welches  $\epsilon$  ist  $P$  kein Outlier?
- (f) Für welches  $\epsilon$  erhält man das Cluster  $\{GHI\}$ ?
- (g) Ist es möglich,  $\epsilon$  so zu wählen, dass man ein einelementiges Cluster erhält?
- (h) Wie viele verschiedene Clusterings können in diesem Dataset gefunden werden?

### Aufgabe 9-2 Outlier Scores

Gegeben sei der folgende 2-dimensionale Datensatz:



Verwenden Sie als Distanzfunktion zwischen den Punkten wieder die Manhattan-Distanz ( $L_1$ -Norm).

Berechnen Sie (unter Ausschluss des Anfragepunktes bei der Berechnung der  $k$ NN):

- Den LOF-Wert für  $k = 2$  für die Punkte  $E$ ,  $K$  und  $O$ .
- Den LOF-Wert für  $k = 4$  für die Punkte  $E$ ,  $K$  und  $O$ .
- Die  $k$ NN-Distanz für  $k = 2$  für alle Punkte.
- Die  $k$ NN-Distanz für  $k = 4$  für alle Punkte.
- Die aggregierten  $k$ NN-Distanzen für  $k = 2$  und  $k = 4$  für alle Punkte.

Diskutieren Sie die Wahl von  $k$  für diesen Datensatz.

Kontrollieren Sie ihre Ergebnisse mit Hilfe von ELKI!