

# MINIMUM SPANNING TREE E CLUSTERING

- SUPPONIAMO DI AVERE UNA COLLEZIONE DI OGGETTI (ES. IMMAGINI, TESTI, RECORD, ECC.) CHE SI INTENDE ORGANIZZARE IN GRUPPI (CLUSTER) DI OGGETTI SIMILI
- E' NECESSARIO ALLORA AVERE UNA **FUNZIONE DISTANZA** TRA GLI OGGETTI IN MODO DA POTER STABILIRE IL GRADO DI SIMILARITA' TRA I VARI OGGETTI  
(ES. NEL CASO DI IMMAGINI DELLA STESSA DIMENSIONE UNA MISURA DI DISTANZA POTREBBE ESSERE IL NUMERO DI PIXEL CORRISPONDENTI CHE DIFFERISCONO OLTRE A UNA CERTA SOGLIA ASSEGNATA)
- IL PROBLEMA DEL **CLUSTERING** E' IL PROBLEMA DI SUDDIVIDERE OPPORTUNAMENTE UN DATO INSIEME DI OGGETTI IN MODO TALE CHE OGGETTI "VICINI" APPARTENGANO AD UNO STESSO CLUSTER E OGGETTI "LONTANI" APPARTENGANO A CLUSTER DIVERSI

## CLUSTERING DI MASSIMA SEPARAZIONE

- UN PROBLEMA DI CLUSTERING CHE UTILIZZA FONDAMENTALMENTE L'ALGORITMO DI KRUSKAL E' QUELLO IN CUI SI CERCA UN CLUSTERING CHE ASSICURI LA MASSIMA DISTANZA POSSIBILE TRA DUE CLUSTER DISTINTI.

- SIA  $U = \{p_1, p_2, \dots, p_n\}$  UN INSIEME DI OGGETTI DOTATO DI UNA FUNZIONE DISTANZA  $d: U \times U \rightarrow \mathbb{R}_0^+$  TALE CHE:

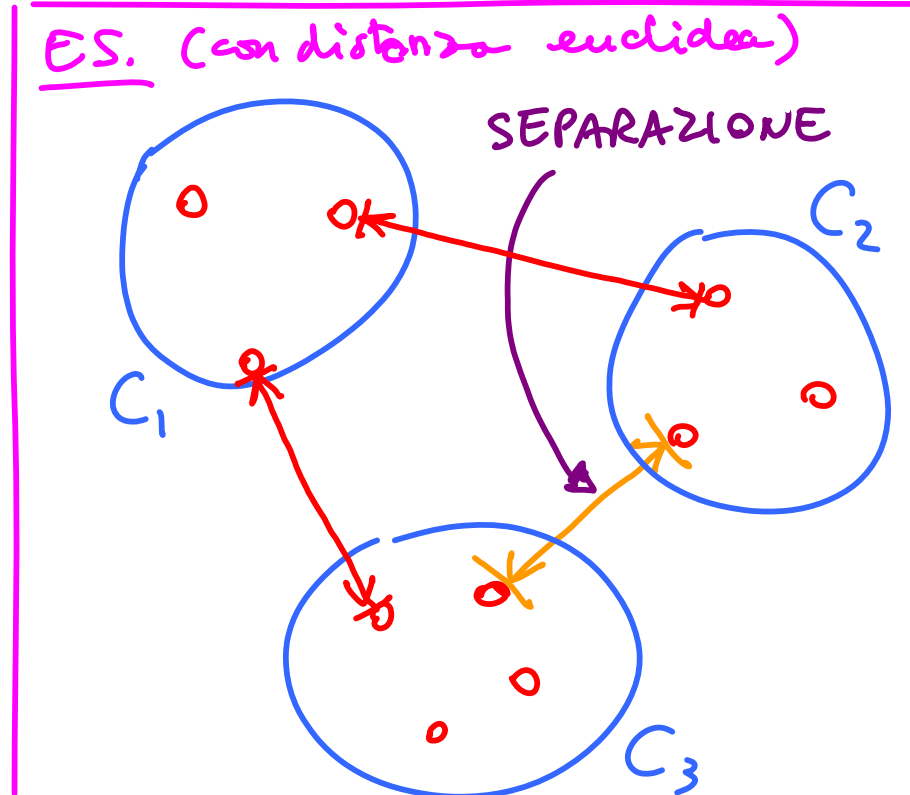
- $d(p_i, p_i) = 0$ , PER OGNI  $i = 1, \dots, n$
- $d(p_i, p_j) > 0$ , PER OGNI  $1 \leq i, j \leq n$ ,  $i \neq j$
- $d(p_i, p_j) = d(p_j, p_i)$ , PER OGNI  $1 \leq i, j \leq n$

(NON E' RICHIESTA LA PROPRIETA' TRIANGOLARE:

$$d(p_i, p_j) \leq d(p_i, p_k) + d(p_k, p_j), \text{ PER OGNI } 1 \leq i, j, k \leq n)$$

- SIA  $k \leq n$  ASSEGNATO (NUMERO DI CLUSTER)
- UNA PARTIZIONE  $C_1, C_2, \dots, C_k$  DI  $U$ , TALE CHE  
 $C_i \neq \emptyset$  PER  $i = 1, \dots, k$ , E' DETTA  
 $k$ -CLUSTERING DI  $U$
- DATO UN  $k$ -CLUSTERING  $\mathcal{C} = \{C_1, C_2, \dots, C_k\}$  DI  $U$ , LA  
 SEPARAZIONE DI  $\mathcal{C}$  E' DATA DA:

$$sp(\mathcal{C}) = \min_{\substack{p \in C_i \\ q \in C_j \\ i \neq j}} d(p, q)$$



- DATI :

- UN INSIEME  $U$  CON FUNZIONE DISTANZA  $d$
- UN INTERO  $1 \leq k \leq |U|$

IL PROBLEMA DEL  $k$ -CLUSTERING A SEPARAZIONE MASSIMA PER  $(U, d)$  E' IL PROBLEMA DI TROVARE UN  $k$ -CLUSTERING  $\mathcal{C}$  DI  $U$  TALE CHE

$$sp(\mathcal{C}) \geq sp(\mathcal{C}'),$$

PER OGNI ALTRO  $k$ -CLUSTERING  $\mathcal{C}'$  DI  $U$

# ALGORITMO KRUSKAL/CLUSTERING $(\{p_1, p_2, \dots, p_m\}, d, k)$

Clustering :=  $\{\{p_1\}, \{p_2\}, \dots, \{p_m\}\}$

- si consideri il grafo completo  $G$  su  $U$
- seguendo un ordinamento non-decrescente degli archi di  $G$  (rispetto alla distanza  $d$ ) e sino a quando  $|\text{Clustering}| = k$  si eseguano le istruzioni:
  - \* sia  $(p_i, p_j)$  l'arco di  $G$  da trattare
  - \* se  $p_i$  e  $p_j$  appartengono a cluster diversi, si uniscano i cluster contenenti  $p_i$  e  $p_j$  in un unico cluster

return Clustering

IN PRATICA, SI TRATTA DELL'ALGORITMO DI KRUSKAL, CON LE SEGUENTI MODIFICHE:

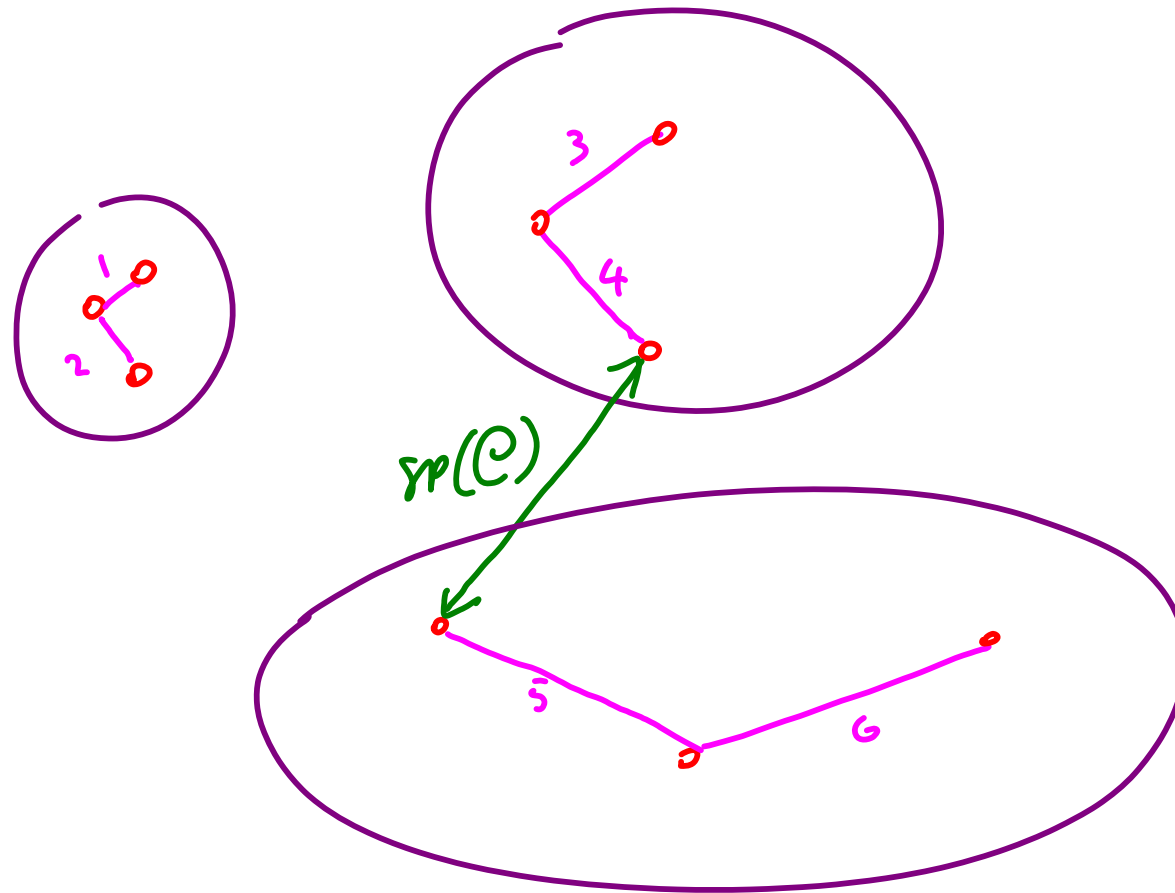
- L'ALGORITMO SI FERMA QUANDO IL NUMERO DI ALBERI NELLA FORESTA DI KRUSKAL E'  $k$  (E NON 1)
- L'ALGORITMO RESTITUISCE LE COMPONENTI CONNESSE DELLA FORESTA TROVATA

COMPLESSITA':  $O(m^2 \lg m)$

OSSERVAZIONE

LA SEPARAZIONE DEL CLUSTERING COSI' CALCOLATO E' MAGGIORE O UGUALE AL PESO  $d(p_i, p_j)$  DELL'ARCO  $(p_i, p_j)$  CHE L'ALGORITMO DI KRUSKAL/CLUSTERING AUREBBE PRESO IN CONSIDERAZIONE ALL'ITERAZIONE SUCCESSIVA SE NON NE FOSSE STATA INTERRUPTA L'ESECUZIONE

ESEMPIO: 3-CLUSTERING (DISTANZA EUCLIDEA)



## CORRETTEZZA

- SIA  $\mathcal{C}^* = \{C_1, C_2, \dots, C_k\}$  IL  $k$ -CLUSTERING COSTRUITO DALL'ALGORITMO DI KRUSKAL/CLUSTERING
- COME OSSERVATO PRIMA,  $sp(\mathcal{C}^*) \geq d^*$ , DOVE  $d^*$  È IL PESO DELL'ARCO CHE L'ALGORITMO DI KRUSKAL/CLUSTERING AVREBBE SELEZIONATO ALL'ITERAZIONE SUCCESSIVA SE NON NE FOSSE STATA INTERRUPTA L'ESECUZIONE (ESSENDO STATO RAGGIUNTO IL NUMERO DI CLUSTER DESIDERATO)
- SIA  $\mathcal{C}' = \{C'_1, C'_2, \dots, C'_k\}$  UN QUALUNQUE ALTRO  $k$ -CLUSTERING DI  $(U, d)$  DISTINTO DA  $\mathcal{C}^*$ .
- OCCORRE VERIFICARE CHE  $sp(\mathcal{C}') \leq sp(\mathcal{C}^*)$ .



- SIA  $C_r \in \mathcal{C}^*$  TALE CHE  $C_r$  INTERSECHI ALMENO DUE CLUSTER DISTINTI DI  $\mathcal{C}'$  (ESISTE NECESSARIAMENTE IN QUANTO  $\mathcal{C}^* \neq \mathcal{C}'$ )

- SIA  $p \in C_r$  E SIA  $C'_s \in \mathcal{C}'$  TALE CHE  $p \in C'_s$

- SIA  $q \in C_r$  TALE CHE  $q \notin C'_s$

- POICHE'  $p, q$  STANNO NELLO STESSO CLUSTER  $C_r$ , ESSI SONO CONNESSI IN  $G$  DA UN CAMMINO  $\pi$  I CUI ARCHI HANNO PESO  $\leq d^*$

- QUINDI SU  $\pi$  VI SONO DUE NODI CONTIGUI  $p', p''$  TALI CHE  $p' \in C'_s, p'' \notin C'_s$ .

MA  $d(p', p'') \leq d^*$  E QUINDI

$sp(\mathcal{C}') \leq d^* \leq sp(\mathcal{C}^*),$  DA CUI LA TESI,

