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 $d : U \times U \to \mathbb{R}_0^+$ TALE CHE:

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

(NON E' RICHIESTA LA PROPRIETA' TRIANGOLARE:

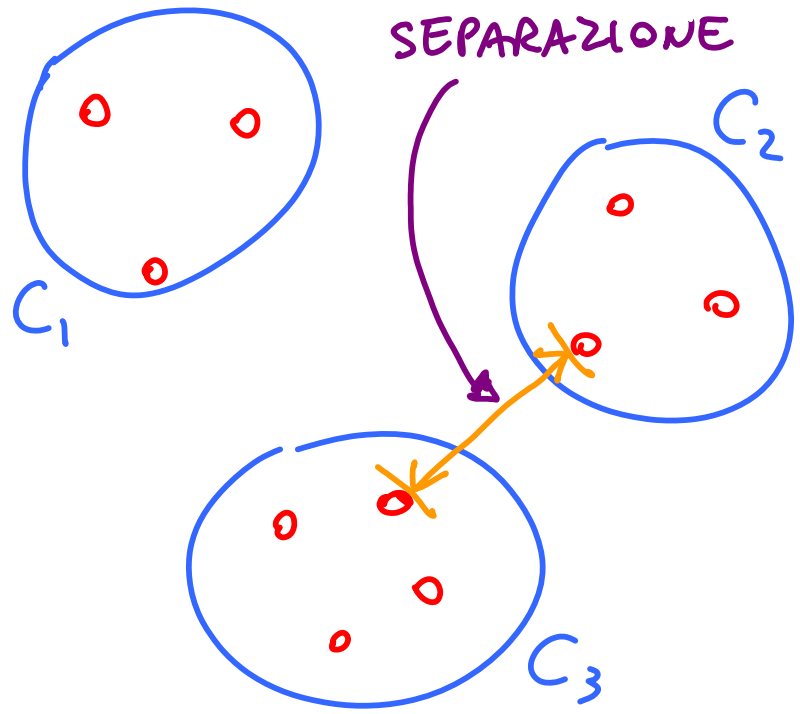
$$d(p_i, p_j) \le d(p_i, p_k) + d(p_k, p_j), \text{ PER OGNI } 1 \le i, j, k \le n$$

)

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

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

ES. (con distanza euclidea)



- 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 $(\{p_1, p_2, \ldots, p_n\}, d, k)$

Clustering := $\{\{p_1\}, \{p_2\}, \ldots, \{p_n\}\}$

- 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

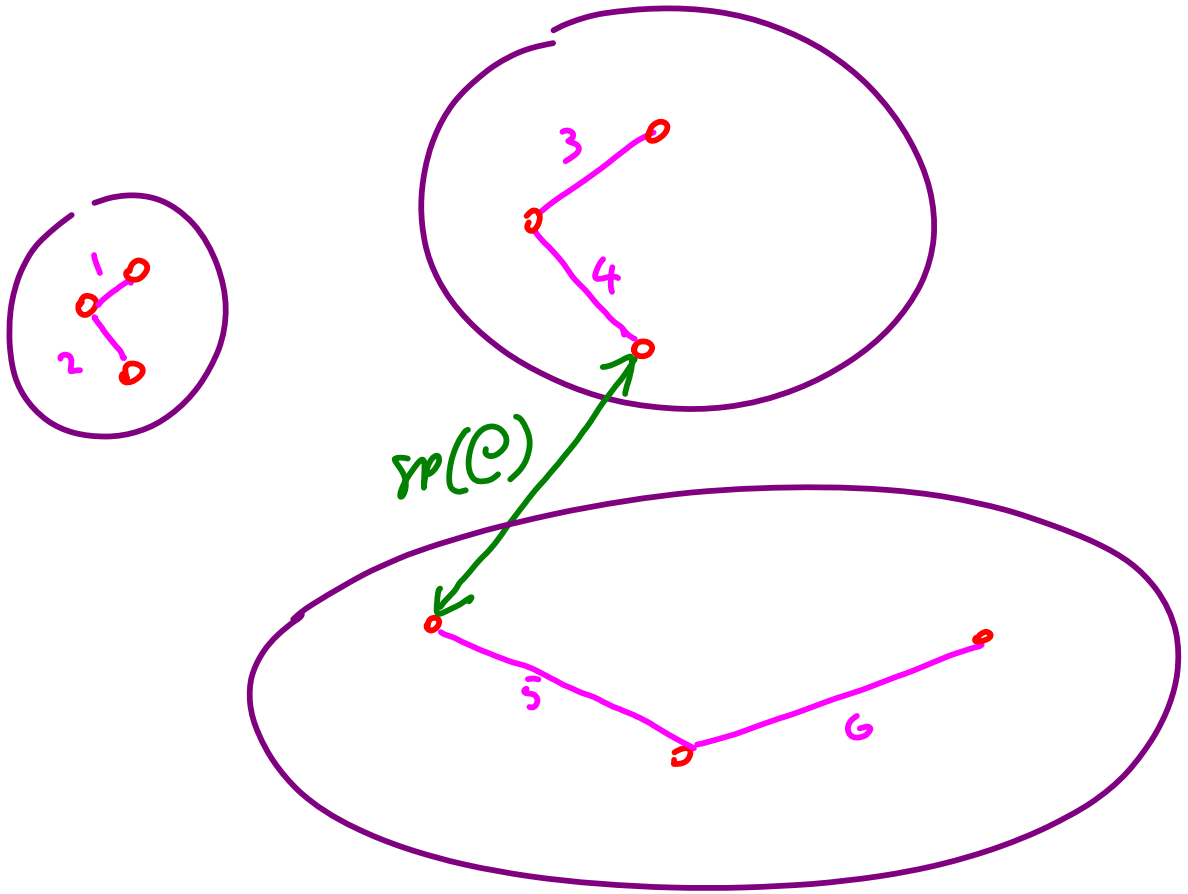
IN PRATICA, SI TRATTA DELL'ALGORITMO DI KRUSKAL, CON LE SEGUENTI MODIFICHE:

- 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': $O(n^2 \lg n)$

OSSERVAZIONE LA SEPARAZIONE DEL CLUSTERING COSÌ CALCOLATO E' UGUALE AL PESO $d(P_i, P_j)$ DELL'ARCO (P_i, P_j) CHE L'ALGORITMO DI KRUSKAL/CLUSTERING AVREBBE UTILIZZATO PER LA SUCCESSIVA UNIONE SE NON NE FOSSE STATA INTERROTTA L'ESECUZIONE

ESEMPIO: 3 - CLUSTERING (DISTANZA EUCLIDEA)



CORRETTEZZA

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

- POICHE' $\mathcal{C}^* \neq \mathcal{C}'$, DEVE NECESSARIAMENTE ESISTERE $C_r \in \mathcal{C}^*$ TALE CHE $C_r \not\subseteq C'_s$, PER OGNI $C'_s \in \mathcal{C}'$.

- QUINDI ESISTONO $p, q \in C_r$ TALI CHE $p \in C'_s$, $q \in C'_t$, CON $C'_s \neq C'_t$

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

- QUINDI SU π ESISTONO DUE NODI CONTIGUI p', p'' TALI CHE $p' \in C'_s$, $p'' \notin C'_s$. MA $d(p', p'') \leq d^*$ E QUINDI $sp(\mathcal{C}') \leq d^*$, COME VOLEVASI DIMOSTRARE.