

Cognome e Nome _____

Matricola _____

Big data – Graph Mining A.A. 2022/2023

Appello del 15 febbraio 2023

Esercizio 1 [8 punti]

Data un grafo non diretto pesato, si definisce *grado pesato* di un nodo i la somma dei pesi degli archi incidenti al nodo i .

Si progettino (eventualmente utilizzando più step di computazione tra loro concatenati) le funzioni *Map* e *Reduce* per restituire in output il **grado pesato massimo** (cioè il grado pesato del nodo avente grado pesato massimo) di un **grafo sociale non diretto pesato** $G=(V, E, w)$, utilizzando spazio locale $O(|V|^{1/2})$. Si assuma che i nodi siano rappresentati da numeri interi, e che un arco del grafo di input, che unisce i nodi i e j ed ha peso x , sia codificato con la tripla (i, j, x) (si assuma, senza perdita di generalità, che $i < j$). Essendo il grafo sociale, si può assumere che il numero di archi incidenti ad ogni nodo sia costante.

Suggerimento: si utilizzi un primo step di computazione per calcolare il grado pesato di ogni nodo; successivamente, si utilizzino gli step di computazione necessari a trovare il massimo tra i gradi pesati calcolati.

Esercizio 2 [9 punti]

Si considerino gli insiemi descritti nella tabella qui sotto.

- **[4 punti]** Si scriva la colonna dell'insieme S_2 in modo che la Jaccard Similarity tra S_1 e S_2 sia $1/2$ e la Jaccard Similarity tra S_2 e S_4 sia $2/5$.
- **[5 punti]** Si calcoli, applicando il metodo visto a lezione ed evidenziando tutti i passaggi fatti riga per riga, la matrice delle firme minhash degli insiemi S_1, S_2, S_3 e S_4 secondo le permutazioni indotte dalle 3 funzioni hash evidenziate in tabella.

Riga	S_1	S_2	S_3	S_4	$(x+2) \bmod 7$	$(2x-1) \bmod 7$	$(3x+2) \bmod 7$
0	0		1	0	2		
1	0		0	1	3		
2	1		0	0	4		
3	1		0	1	5		
4	1		1	1	6		
5	0		0	0	0		
6	0		1	1	1		

Esercizio 3 [8 punti]

Si consideri l'algoritmo *CURE* per il clustering di punti in uno spazio euclideo con d dimensioni.

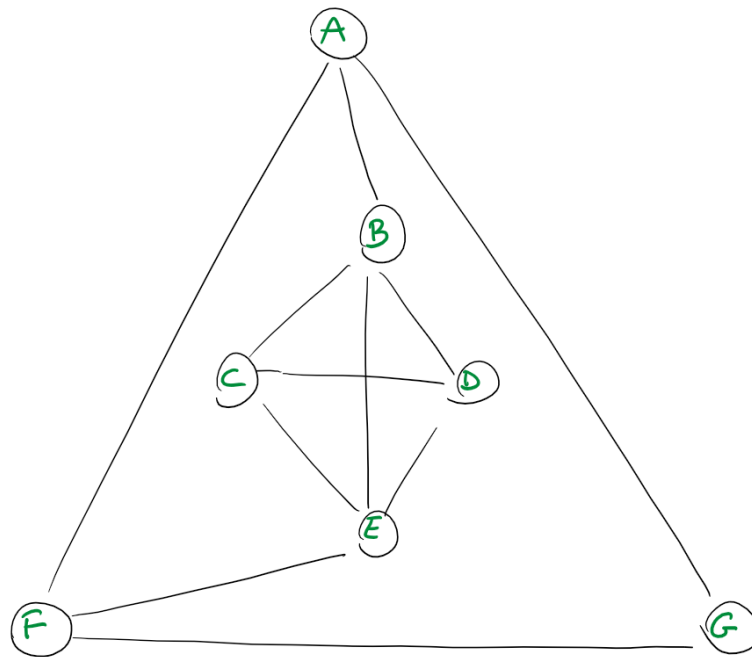
Si dica in modo **sintetico e puntuale**, giustificando le risposte:

- Quali sono le assunzioni che tale algoritmo fa circa la conformazione dei cluster.
- Come viene identificato e memorizzato un cluster
- Se tale algoritmo è di tipo gerarchico e/o di tipo con assegnamento di punti

Infine, di descriva l'algoritmo evidenziando le sue varie fasi.

Esercizio 4 [9 punti]

Si consideri la rete sociale in figura.



Si approssimi la **edge betweenness** (con l'algoritmo di **Girvan–Newman**) assumendo di considerare solo i contributi dovuti alle visite radicate ai nodi **a**, **b**, **c** e **g**.

Si discuta se tale calcolo può portare a dividere il grafo in due o più cluster, giustificando la risposta.

Regole per lo svolgimento della prova scritta:

- Per svolgere il compito si hanno a disposizione **90** minuti
- Scrivere **subito** nome, cognome, matricola su **OGNI FOGLIO (compreso questo)**.
- Durante la prova scritta **non** è possibile abbandonare l'aula.
- Non è ammesso **per nessun motivo** comunicare in qualsiasi modo con altre persone
- Non è possibile consultare appunti, libri e dispense.
- Qualsiasi strumento elettronico di calcolo o comunicazione (telefoni cellulari, calcolatrici, palmari, computer, etc...) deve essere **completamente disattivato** e **depositato in vista sulla cattedra**
- Mettere in vista sul banco un valido documento di identità.