Cognome e Nome _____

Matricola _____

Big data – Graph Mining
A.A. 2023/2024

Appello dell'8 febbraio 2024

Esercizio 1 [8 punti]

Data un grafo diretto pesato, si definisce *grado entrante pesato* di un nodo *i* la somma dei pesi degli archi entranti nel nodo *i (aventi cioè la freccia diretta verso il nodo i)*.

Si progettino (eventualmente utilizzando più step di computazione tra loro concatenati) le funzioni *Map* e *Reduce* per restituire in output ***tutti i nodi avente grado entrante pesato massimo*** di un **grafo sociale diretto pesato** G=(V, E, w), utilizzando spazio locale $O(|V|^{1/3})$. Si assuma che i nodi siano rappresentati da numeri interi, e che un arco del grafo di input, che va dal nodo *i* al nodo *j* ed ha peso x, sia codificato con la tripla (i, j, x). 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 entrante pesato di ogni nodo; successivamente, si utilizzino gli step di computrazione necessari a trovare i nodi con il massim grado entrante.

Esercizio 2 [9 punti]

- **[1,5 punti]** Si definisca la Jaccard similarity di due insiemi.
- **[2,5 punti]** Dato il dominio {0, 1, 2, 3, 4, 5, 6}, si forniscano tre insiemi **S_1, S_2, S_3** tali che la Jaccard Similarity di S_1 e S_2 sia 1/2 e la Jaccard Similarity di S_2 e S_3 sia 1/3. Si calcoli inoltre la Jaccard Similarity di S_1 e S_3. Si riportino gli insiemi in tabella.
- **[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 e S_3 secondo le permutazioni indotte dalle 3 funzioni hash evidenziate in tabella.

Riga	S_1	S_2	S_3	(x+1) mod 7	(x-1) mod 7	(2x+1) mod 7
0						
1						
2						
3						
4						
5						
6						

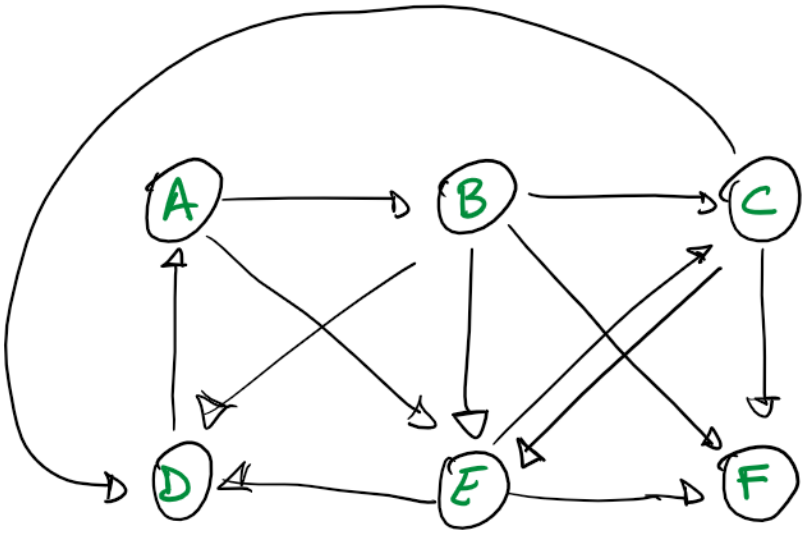
Esercizio 3 [8 punti]

Si consideri l'algoritmo di *Bradley, Fayyad e Reina* per il clustering di punti in uno spazio euclideo con d dimensioni.
Si dica in modo **sintetico e puntuale, giustificando la risposta**:

a. Se tale algoritmo è di tipo gerarchico o di tipo con assegnamento di punti
b. L'assunzione di fondo che questo algoritmo fa sui punti di ogni cluster
c. Cosa sono i *discard set, i compressed set e i retained set* e perché vengono chiamati così.
d. Come è possibile ricostruire le informazioni su media e varianza dei cluster a partire dalle informazioni memorizzate per ogni cluster.

Esercizio 4 [9 punti]

Si consideri la rete sociale in figura.
Applicando l'algoritmo A-priori, **si individuino i cluster indotti da soglia di support uguale almeno a 2**.
Si mostri ogni passo della computazione effettuata.



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.