Cognome e Nome _____

Matricola _____

Big data – Graph Mining
A.A. 2024/2025

Appello del 15 gennaio 2025

PRIMA PARTE

Esercizio 1 [8 punti]

Siano date in input:

- Delle triple **(titolo, lista_id_autori, copie_vendute)** che rappresentano i dati di un libro che ha venduto **copie_vendute** copie ed è stato scritto dagli autori presenti nella lista **lista_id_autori**.
- Delle coppie **(id_autore, nome_cognome)** che contengono le informazioni circa il nome e cognome dell'autore con id **id_autore** (che varia da 0 a n-1).

Si progettino (eventualmente utilizzando più step di computazione tra loro concatenati) le funzioni *Map e Reduce* per dare in output **nome_cognome** di uno degli autori che hanno venduto complessivamente più copie, utilizzando spazio locale $O(n^{1/2})$.
Si discuta il *reducer size* di tutti gli step individuati, assumendo che ogni autore ha scritto un numero costante di libri (non dipendente quindi da *n*).

Suggerimento: prima trovare, per ogni autore, il numero di copie vendute (ottenendo così delle triple **(id_autore, nome_cognome, totale_copie_vendute)**), e poi calcolare un autore che ha venduto il più grande numero di copie.

Esercizio alternativo [valutazione massima 4 punti]: assumere di avere in input le triple **(id_autore, nome_cognome, totale_copie_vendute)** (con id_autore che varia da 0 a n-1) e dare in output **nome_cognome** di uno degli autori che hanno venduto complessivamente più copie, utilizzando spazio locale $O(n^{1/2})$.

Esercizio 2 [9 punti]

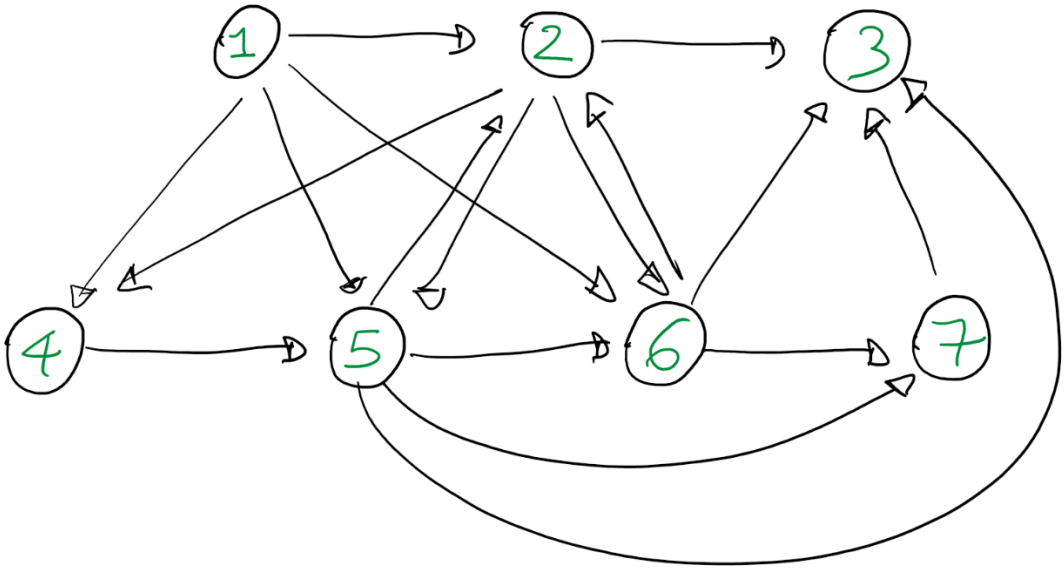
- **[1,5 punti]** Si definisca la Jaccard similarity di due insiemi.
- **[2,5 punti]** Dato il dominio {0, 1, 2, 3, 4, 5, 6} e l'insieme $\mathbf{S_2}$={1, 2, 3, 6}, si forniscano gli insiemi $\mathbf{S_1}$, $\mathbf{S_3}$ tali che la Jaccard Similarity di S_1 e S_2 sia 1/2 e la Jaccard Similarity di S_2 e S_3 sia 2/5. 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+2) mod 7	(x-1) mod 7	(2x+3) mod 7
0		**0**				
1		**1**				
2		**1**				
3		**1**				
4		**0**				
5		**0**				
6		**1**				

SECONDA PARTE

Esercizio 3 [9 punti]

Si consideri la rete sociale in figura.



Applicando l'algoritmo A-priori, si individuino i cluster di almeno 5 nodi indotti da soglia di support uguale almeno a 2. Si mostri ogni passo della computazione effettuata.

Esercizio 4 [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. L'assunzione di fondo che questo algoritmo fa sui punti di ogni cluster
b. Cosa sono i *discard set, i compressed set e i retained set* e perché vengono chiamati così.
c. Come è possibile ricostruire le informazioni su media e varianza dei cluster a partire dalle informazioni memorizzate per ogni cluster.
d. Come è definita la distanza di Mahalanobis e che ruolo ha nell'algoritmo BFR

Regole per lo svolgimento della prova scritta:

- Per svolgere il compito si hanno a disposizione **45** minuti per la seconda parte e **90** minuti per l'intero compito. Chi non consegna entro i primi 45 minuti automaticamente rinuncia al voto dell'esonero.
- 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 lontano dalla propria postazione.**
- Mettere in vista sul banco un valido documento di identità.

Esempio di input per Esercizio 1:

- (I Promessi Sposi, (0), 312000)
- (Acqua in bocca, (1,2), 10000)
- (La forma dell'acqua, (1), 8000)
- (0, Alessandro Manzoni)
- (1, Andrea Camilleri)
- (2, Carlo Lucarelli)

In questo input abbiamo 3 libri (uno per ogni tripla) e 3 autori (uno per ogni coppia).

Il suggerimento consiglia di calcolare prima le triple
(id_autore, nome_cognome, totale_copie_vendute)
che nell'esempio risulteranno essere:

- (0, Alessandro Manzoni, 312000)
- (1, Andrea Camilleri, 18000)
- (2, Carlo Lucarelli, 10000)

Per poi calcolare l'autore che ha venduto il maggior numero di copie (Alessandro Manzoni nell'esempio).