Cognome e Nome _____

Matricola _____

Big data – Graph Mining
A.A. 2025/2026

Appello del 22 gennaio 2026

Esercizio 1 [8 punti]

Sia dato un grafo sociale diretto che modella relazioni di *simpatia/antipatia* tra persone. Un nodo corrisponde ad una persona. Gli archi sono di due tipi: un arco di *simpatia* (rispettivamente *antipatia*) diretto dal nodo i al nodo j indica che la persona j è ritenuta *simpatica* (risp. *antipatica*) dalla persona i.

Il **punteggio sociale** di una persona i è dato dal numero di archi di simpatia entranti nel nodo i **meno** il numero di archi di antipatia entranti nel nodo i.

Si progettino (eventualmente utilizzando più step di computazione tra loro concatenati) le funzioni *Map* e *Reduce* per restituire in output ***tutte le persone (ovvero i nodi) che hanno il massimo punteggio sociale*** dato un **grafo sociale diretto**, utilizzando spazio locale $O(n^{1/3})$, dove n è il numero di nodi del grafo. Si assuma che i nodi siano rappresentati da numeri interi da 0 a n-1, che un arco di simpatia del grafo, che va dal nodo ***i*** al nodo ***j***, sia codificato dalla tripla **("simpatia", i, j)** e che un arco di antipatia del grafo, che va dal nodo ***i*** al nodo ***j***, sia codificato dalla tripla **("antipatia", i, j)**. Si assuma anche che ogni nodo abbia un numero costante di archi uscenti e di archi entranti.

Suggerimento: si utilizzi un primo step di computazione per calcolare il punteggio sociale di ogni nodo; successivamente, si utilizzino gli step di computrazione necessari a trovare i nodi che realizzano il massimo punteggio sociale.

Esempio di input:

- ("simpatia", 0, 1)
- ("simpatia", 0, 2)
- ("simpatia", 2 ,1)
- ("simpatia", 2, 0)
- ("antipatia", 0, 3)
- ("antipatia", 3, 0)
- ("antipatia", 3, 1)

In questo input abbiamo 4 nodi e 7 archi (4 archi di simpatia e 3 archi di antipatia).

Il suggerimento consiglia di calcolare prima le triple **(id_nodo, punteggio_sociale)**
che nell'esempio risulteranno essere:

- (0, 0)
- (1, 1)
- (2, 1)
- (3, -1)

Per poi calcolare i nodi che realizzano il massimo punteggio sociale (i nodi **1 e 2**).

Esercizio alternativo [valutazione massima 4 punti]: assumere di avere in input le coppie **(id_nodo, punteggio_sociale)** (con id_nodo che varia da 0 a n-1) e dare in output *tutte le persone (ovvero i nodi) che hanno il massimo punteggio sociale*, utilizzando spazio locale $O(n^{1/3})$.

Esercizio 2 (9 punti)

[5 punti] Si consideri la variante multi-hash dell'algoritmo *Park Chen Yu* (con due tabelle hash) nell'ambito del problema degli *itemset frequenti*, e se ne descriva il funzionamento in modo dettagliato.

Si dica inoltre, in modo **sintetico e puntuale, giustificando la risposta**:

a. **[2 punti]** Quando, in questa variante, si può escludere che una coppia sia candidata ad essere frequente.
b. **[2 punti]** Perché viene impiegato il metodo delle triple per memorizzare i contatori delle coppie di itemset candidate ad essere frequenti.

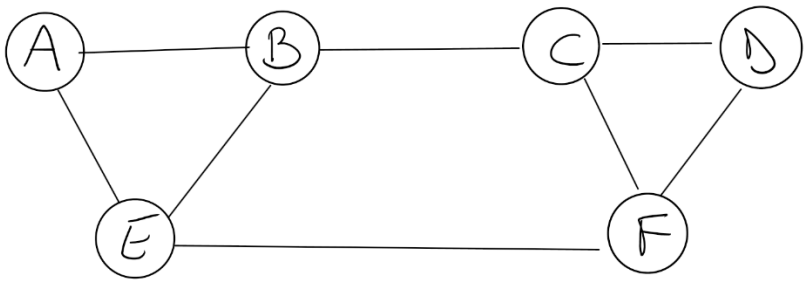
Esercizio 3 [9 punti]

a. **[4 punti]** Si dica in modo **sintetico e puntuale, giustificando la risposta**, che cosa è il PageRank e come/perché è possibile calcolarlo in modo efficiente con il *power iteration method*, spiegando in modo particolare il ruolo del parametro **β** di teleport e cosa accadrebbe quando, in un grafo che presenta uno spider trap e non presenta dead-end, si sceglie **β=1**.

b. **[5 punti]** Si faccia un esempio di grafo dei link del web con **6** nodi e **12** archi, in cui siano presenti esattamente un dead-end ed uno spider trap, e si impostino le equazioni per il calcolo del **Topic-Specific PageRank** con parametro **β** generico, tenendo conto del fatto che l'insieme di teleport deve costituito da **3** nodi a vostra scelta. Si dica in modi chiaro quale è il dead-end, da quali nodi è costituito lo spider trap e quali sono i nodi scelti per l'insieme di teleport.

Esercizio 4 [8 punti]

a. **[6 punti]** Si consideri la rete sociale in figura.



Si calcoli la **edge betweeness** (prima iterazione dell'algoritmo di **Girvan–Newman**) di tutti gli archi.
Si dica se tale calcolo può portare a dividere il grafo in due o più cluster, o se è necessario effettuare ulteriori iterazioni dell'algoritmo di Girvan-Newman, giustificando la risposta.

b. **[2 punti]** Qual è la complessità computazionale dell'algoritmo di Girvan-Newman? Come può essere diminuita?

Regole per lo svolgimento della prova scritta:

- Per svolgere il compito si hanno a disposizione **90** minuti (**70** minuti se non si svolge l'Esercizio 1)
- 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, tablet, computer, smartwatch, etc…) deve essere **completamente disattivato** e **depositato in vista sulla cattedra**
- Mettere in vista sul banco un valido documento di identità.