Cognome e Nome _____

Matricola _____

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

Appello del 4 luglio 2024

Esercizio 1 [8 punti]

Si progettino (eventualmente utilizzando più step di computazione tra loro concatenati) le funzioni *Map* e *Reduce* per generare tutti **i cicli di lunghezza 4 con il loro relativo peso** in un **grafo diretto pesato** G=(V,E,w). Si assuma che un arco del grafo di input, che va dal nodo i al nodo j ed ha peso x, sia codificato come la tripla (i, j, x) e che il peso di un ciclo sia dato dalla somma dei pesi degli archi che lo compongono.

Si chiede di dare in output i cicli **senza ripetizioni**: ad esempio, il ciclo formato dai nodi 2→ 4→3 →10 non dovrà comparire anche sottoforma di 4→3→10→2, 3→ 10→ 2→ 4 o 10→ 2→ 4→ 3.

Suggerimento: prima trovare i path di 2 archi (quindi con 3 nodi) con la tecnica vista a lezione e sucessivamente trovare i cicli di 4 archi analizzando come si possono concatenare i path di 2 archi trovati al primo step.

Esercizio 2 [9 punti]

Si consideri la variante multi-hash con due tabelle hash dell'algoritmo *Park Chen Yu*, nell'ambito del problema degli *itemset frequenti*, e se ne descriva in dettaglio il funzionamento.

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

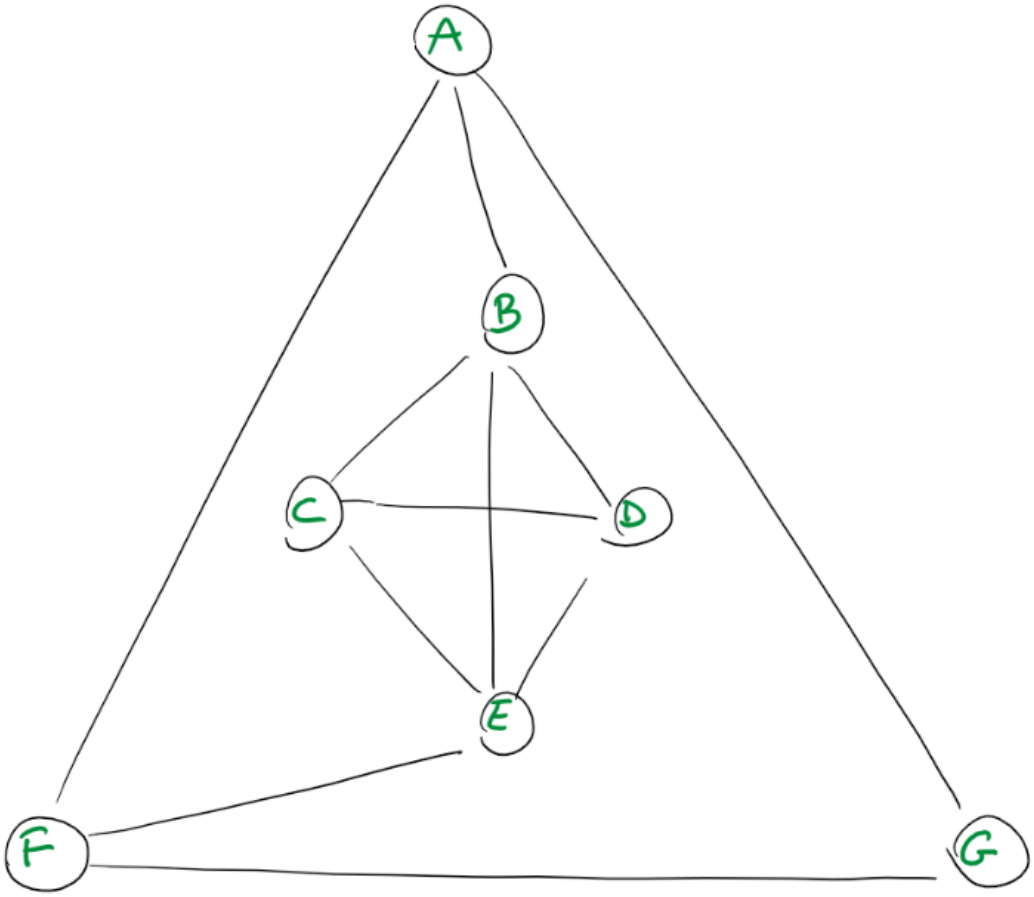
a. cosa si può affermare quando un contatore di una delle due tabelle hash raggiunge la soglia di frequenza;
b. cosa deve accadere perché una coppia venga considerata non candidata ad essere frequente;
c. come avviene la compressione delle tabelle hash;
d. quale vantaggio offre rispetto all'algoritmo PCY con singola tabella hash.

Esercizio 3 [9 punti]

a. Si dica in modo **sintetico e puntuale, giustificando la risposta**, che cosa è il PageRank e come è possibile calcolarlo in modo efficiente, spiegando in modo particolare il ruolo parametro **β** di teleport e cosa accadrebbe quando, in assenza di dead-end, si sceglie **β=1**.
b. Si faccia un esempio di grafo del web con **5** nodi e **9** archi, in cui sia presente un dead-end, 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 due nodi a vostra scelta.

Esercizio 4 [8 punti]

Si consideri la rete sociale in figura.



Si approssimi la **edge betweeness** (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.