

Cognome e Nome _____

Matricola _____

Appello dell'11 settembre 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 **dal massimo dei pesi** degli archi che lo compongono.

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

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

Esercizio 2 (10 punti)

- [2 punti]** Si definisca cosa è il **supporto** di un itemset e quando un itemset può considerarsi **frequente**; si enunci la proprietà della **monotonia della frequenza**.
- [5 punti]** Far vedere sulla seguente istanza l'esecuzione passo passo dell'algoritmo A-priori per individuare tutti gli insiemi di item che abbiano supporto almeno uguale a **3**.
- [3 punti]** Si definisca il bordo negativo e lo si calcoli sulla stessa istanza

basket	item
1	1, 2, 3, 4
2	5, 6, 7, 8, 9, 10
3	1, 3, 5, 6, 7, 8
4	2, 7, 8, 9
5	7, 9
6	1, 5, 6, 7
7	3, 5, 7, 9, 10
8	7, 8, 9
9	5, 7, 8
10	2, 4, 5

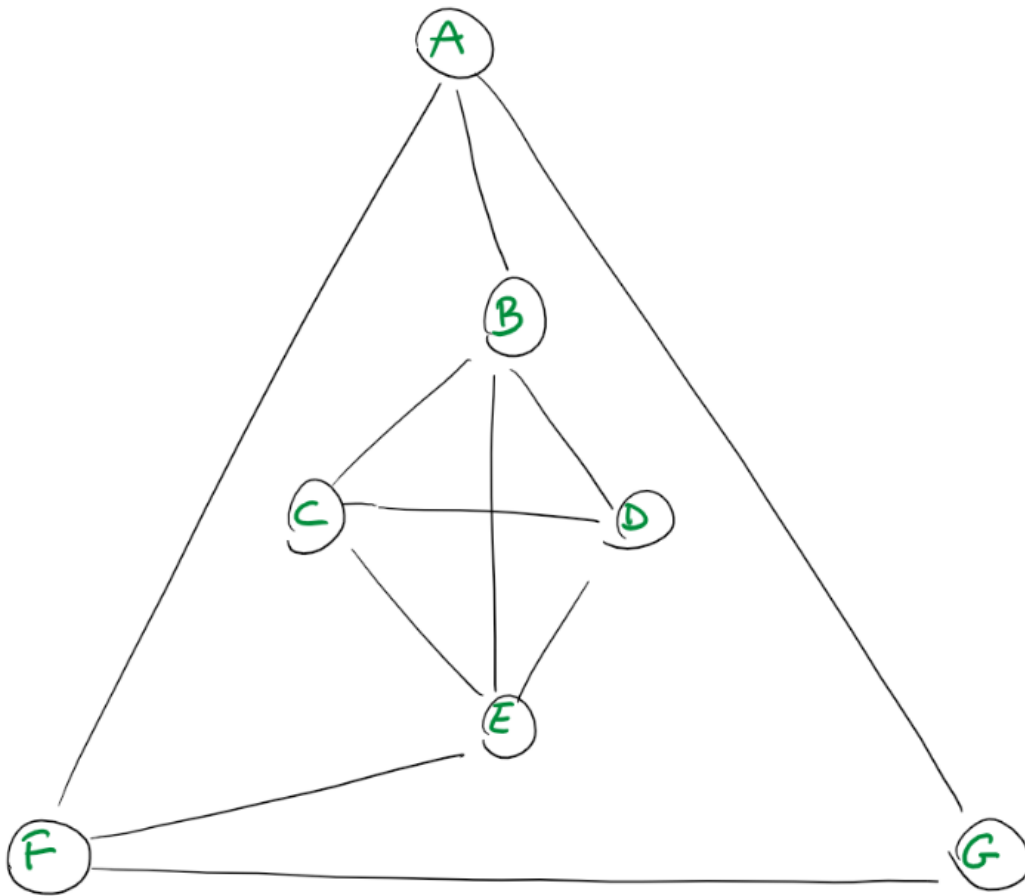
Esercizio 3 [8 punti]

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**:

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

Esercizio 4 [8 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 **d**, **e**, **f** 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.