

Big data – Graph Mining

A.A. 2021/2022

Appello dell'8 giugno 2022

Esercizio 1 (9 punti)

Si progettino (eventualmente utilizzando più step di computazione tra loro concatenati) le funzioni *Map* e *Reduce* per generare tutti i path di lunghezza 4 archi (quindi con 5 nodi) in un **grafo diretto non pesato** $G=(V,E)$. Si assuma che un arco del grafo di input, che va dal nodo i al nodo j , sia codificato come la coppia (i, j) .

Suggerimento: prima trovare i path di 2 archi (quindi con 3 nodi) con la tecnica vista a lezione: il Map genera per ogni arco (i, j) 2 coppie $(j, ("primo", i))$ e $(i, ("secondo", j))$ e il reduce analizza le coppie che hanno in comune la stessa chiave k e...

Esercizio 2 [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 viene memorizzato in memoria un cluster, quanta memoria occupa tale memorizzazione e perché si memorizzano tali informazioni
-

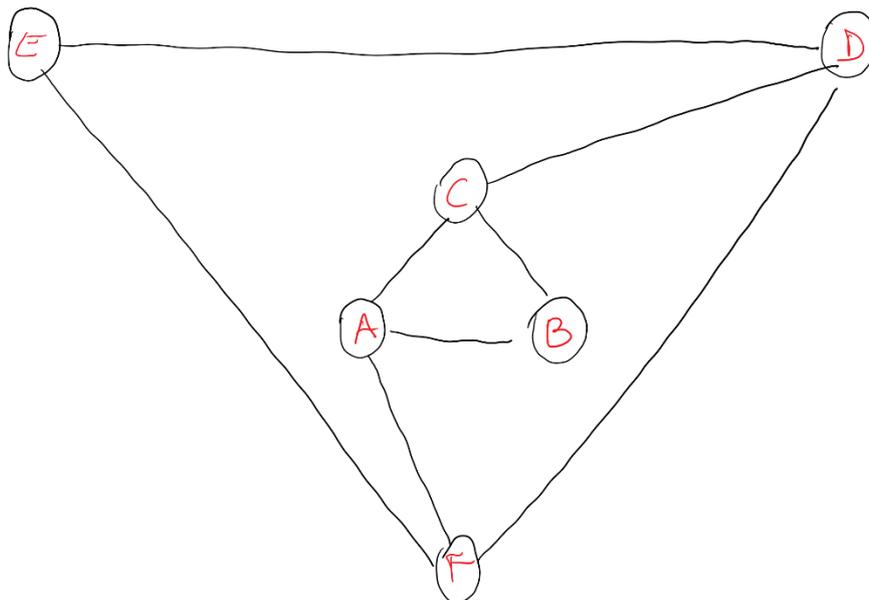
Esercizio 3 [9 punti]

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

- Che cosa è il PageRank e come è possibile calcolarlo, spiegando in modo particolare il ruolo del parametro β di teleport.
- Cosa è il Topic-Specific PageRank e come è possibile calcolarlo
- Che significato ha il SimRank nel contesto delle reti sociali k -partite, e come si applicano le stesse tecniche del PageRank per calcolarlo.

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 **a**, **b**, **e** ed **f**.

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 **100** minuti
- Scrivere **subito** nome, cognome, matricola su OGNI FOGLIO.
- 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.