Big data – Graph Mining
A.A. 2019/2020

Appello del 20 febbraio 2020

Esercizio 1 [9 punti]

Si progettino (eventualmente utilizzando più step di computazione tra loro concatenati) le funzioni *Map* e *Reduce* per calcolare il peso dell'arco con peso più alto in un grafo sociale non diretto G con n nodi e m archi, con grado massimo costante. In particolare, si assuma che in G ogni nodo **i** è associato ad una persona con età $\mathbf{a_i}$ e che il peso di un arco (che denota l'età media della relazione di amicizia) è dato dalla media aritmetica delle età dei suoi estremi.

L'input è dato da n triple **("nodo", i, a_i)** con i=1,…,n per i vertici del grafo sociale G, e da m triple **(k, i , j)** per ogni arco (i, j) del grafo sociale, con k=1,…,m.

L'algoritmo deve usare memoria locale $O(m^{1/2})$ e usare un numero di step di computazione costante (che non dipende da n).

Esercizio alternativo: **[dà luogo a una valutazione massima di 5 punti]** Si assuma che l'input già consista delle coppie **(k, w_k),** per ogni arco del grafo sociale, con k=1,…,m e dove $\mathbf{w_k}$ è il peso del k-esimo arco.

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**:

a. Se tale algoritmo è di tipo gerarchico o di tipo con assegnamento di punti
b. L'assunzione di fondo che questo algoritmo fa sui punti di ogni cluster
c. Come viene memorizzato in memoria un cluster, quanta memoria occupa tale memorizzazione e perché si memorizzano tali informazioni
d. Cosa sono i *discard set, i compressed set e i retained set* e perché vengono chiamati così.
e. Come è definita la distanza di Mahalanobis e che ruolo riveste nell'algoritmo.

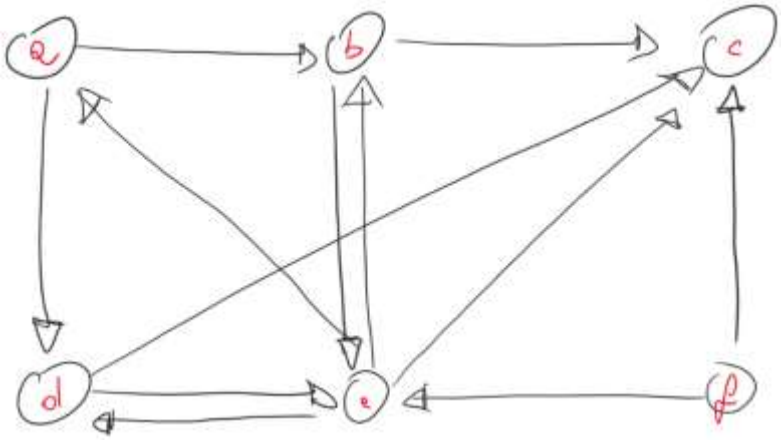
Esercizio 3 [9 punti]

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

a. Che cosa è il PageRank e come è possibile calcolarlo, spiegando in modo particolare il ruolo del parametro **β** di teleport.
b. Cosa è il Topic-Specific PageRank e come è possibile calcolarlo
c. 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.



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

Regole per lo svolgimento della prova scritta:

- Per svolgere il compito si hanno a disposizione **110** 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.