Cognome e Nome _____

Matricola _____

Big data – Graph Mining
A.A. 2022/2023

Appello straordinario dell'8 maggio 2023

Esercizio 1 [8 punti]

Data un grafo diretto, si definisce *grado entrante* di un nodo *i* il numero degli archi entranti nel nodo *i*.

Si progettino (eventualmente utilizzando più step di computazione tra loro concatenati) le funzioni *Map* e *Reduce* per restituire in output il ***grado entrante massimo*** (cioè il numero degli archi entranti del nodo avente il maggior numero di archi entranti) di un **grafo sociale diretto** G=(V, E), utilizzando spazio locale $O(|V|^{1/3})$. Si assuma che i nodi siano rappresentati da numeri interi, e che un arco del grafo di input, che va dal nodo *i* al nodo *j*, sia codificato con la coppia (i, j). Essendo G un grafo sociale, si può assumere che il numero di archi entranti in ogni suo nodo sia costante.

Suggerimento: si utilizzi un primo step di computazione per calcolare il grado entrante di ogni nodo; successivamente, si utilizzino gli step di computrazione necessari a trovare il massimo tra i gradi entranti calcolati.

Esercizio 2 [9 punti]

Si consideri il problema dei Frequent Itemset.

a. Si definisca il bordo negativo e si indichi come è possibile calcolarlo
b. Si descriva il funzionamento dell'algoritmo di Tovoinen
c. Si definiscano i falsi negativi e i falsi positivi, e si discuta come vengono gestiti dall'algoritmo di Tovoinen

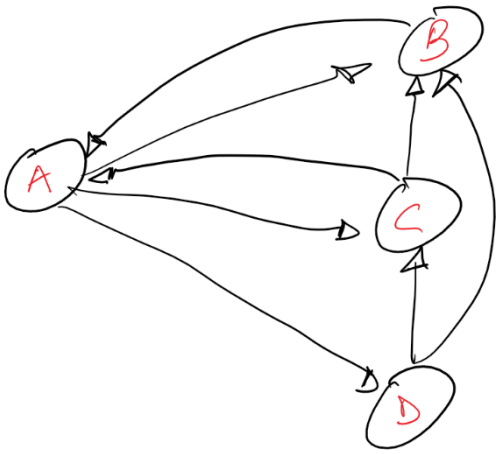
Esercizio 3 [8 punti]

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**:

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. Cosa sono i *discard set, i compressed set e i retained set* e perché vengono chiamati così.
d. Come viene memorizzato in memoria un cluster, quanta memoria occupa tale memorizzazione e perché si memorizzano tali informazioni

Esercizio 4 [9 punti]

Si consideri la rete in figura.



a. Scrivere le equazioni del sistema per il calcolo del PageRank, assumendo un parametro **β** di teleport generico.
b. Calcolare come cambia il PageRank dopo una iterazione, a partire dai valori $p_A=p_B=0.3$ e $p_C= p_D=0.2$ (considerando **β**=0.9, e assumendo quindi che con probabilità 1- **β**=0.1 un random surfer salti verso una pagina a caso).
c. Si riscrivano le equazioni del punto a. nel caso in cui si voglia calcolare il topic-specific PageRank con insieme di teleport uguale a **{C, D}**.

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.