

Cognome e Nome _____

Matricola _____

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

Appello del 7 giugno 2023

Esercizio 1 [8 punti]

Dato un grafo non diretto, si definisce *grado* di un nodo i il numero degli archi incidenti al nodo i .

Si progettino (eventualmente utilizzando più step di computazione tra loro concatenati) le funzioni *Map* e *Reduce* per restituire in output il **grado massimo** (cioè il numero degli archi incidenti al nodo avente il maggior numero di archi incidenti) di un **grafo sociale non 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 unisce i nodi i e j , sia codificato con la coppia (i, j) (si assuma, senza perdita di generalità, che $i < j$). Essendo G un grafo sociale, si può assumere che il numero di archi incidenti ad ogni suo nodo sia costante (non dipendente da $|V|$).

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

Esercizio 2 [9 punti]

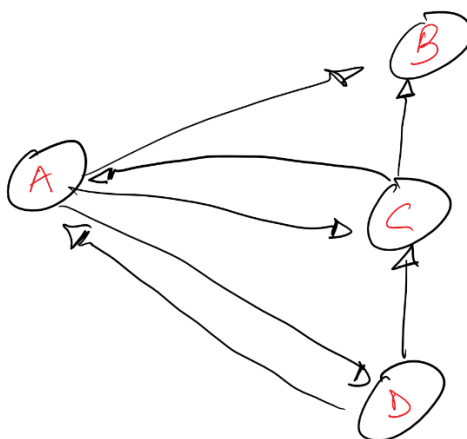
Si consideri l'algoritmo *Park Chen Yu* (con singola tabella hash) nell'ambito del problema degli *itemset frequenti*, e se ne descriva in dettaglio il funzionamento.

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

- cosa si può affermare quando un contatore della tabella hash raggiunge la soglia di frequenza, e cosa si può dire quando non la raggiunge;
- come operano le varianti multi-stage e multi-hash dell'algoritmo PCY, e quali sono i loro vantaggi e svantaggi.

Esercizio 3 [9 punti]

Si consideri la rete in figura (si noti che il nodo B è un dead end).



- Scrivere le equazioni del sistema per il calcolo del PageRank, assumendo un parametro β di teleport generico.
- Calcolare come cambia il PageRank dopo una iterazione, a partire dai valori $p_A=p_B=0.4$ e $p_C=p_D=0.1$ (considerando $\beta=0.9$, e assumendo quindi che con probabilità $1-\beta=0.1$ un random surfer salti verso una pagina a caso).
- Si riscrivano le equazioni del punto a. nel caso in cui si voglia calcolare il topic-specific PageRank con insieme di teleport uguale a $\{A, D\}$.

Esercizio 4 [8 punti]

Si descriva come operano gli algoritmi di clustering gerarchici o agglomerativi che partono da tutti cluster singoletti.

Si dica in modo **sintetico e puntuale**, giustificando le risposte:

- in che modo possono essere scelti i cluster da unire;
- quali possono essere le condizioni di terminazione del clustering.

Infine, si faccia un esempio di clustering sul piano euclideo (partendo da un insieme di almeno 8 punti) applicando un criterio di unione di cluster a scelta e andando avanti fino ad ottenere un unico cluster. Si mostri anche il dendrogramma (l'albero di unione dei cluster) corrispondente al processo di clusterizzazione effettuato.

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.