Big data – Graph Mining
A.A. 2021/2022

Appello del 15 settembre 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 cicli di lunghezza 3** 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…

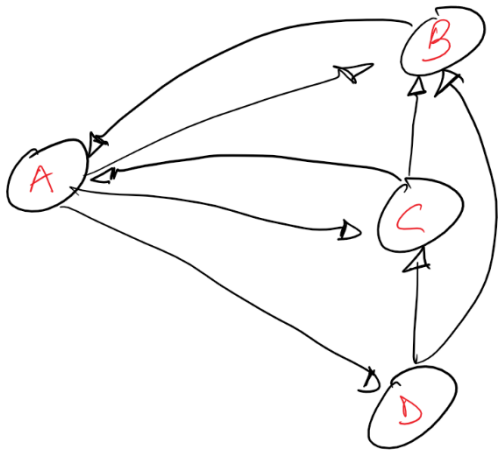
Esercizio 2 (8 punti)

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**:

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.

Esercizio 3 [9 punti]

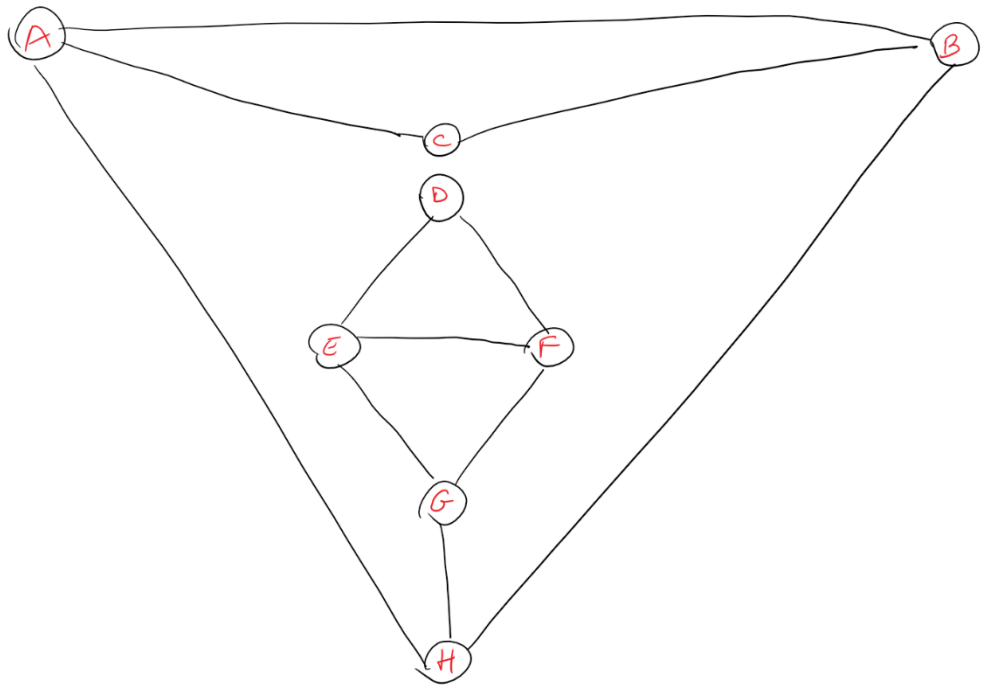
Si consideri la rete in figura.



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=0.4 e p_B= p_C= p_D=0.2 (considerando **β**=0.85, e assumendo quindi che con probabilità 1- **β**=0.15 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 **{A, B}**.

Esercizio 4 [8 punti]

Si consideri la rete sociale in figura.



Si approssimi la **edge betweeness** (con l'algoritmo di Girvan–Newman) assumendo di considerare solo i contributi dovuti alle visite radicate ai nodi **a**, **g**, **f** ed **h**.

Si discuta se e come tale calcolo può portare a dividere il grafo in 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.