Cognome e Nome _____

Matricola _____

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

Appello del 23 gennaio 2020

PRIMA PARTE

Esercizio 1 (14 punti)

Si progettino (eventualmente utilizzando più step di computazione tra loro concatenati) le funzioni *Map* e *Reduce* per calcolare il **raggio** di un cluster C di n punti in uno spazio metrico con funzione di distanza d, definito come la massima distanza tra il *centroide* **c** e un *punto del cluster*.
Si assuma che il centroide sia già noto in quanto precedentemente calcolato.
L'input è dato dalle coppie **(i, x_i)** con i=1,…,n per i punti del cluster C, e il centroide **c** è dato nella coppia **(0, c).**
L'algoritmo deve usare memoria locale $O(n^{1/3})$ e usare un numero di step di computazione costante (che non dipende da n).

Esercizio 2 (10 punti)

Si considerino le seguenti stringhe che rappresentano insieme di item (a lettere uguali corrisponde lo stesso item), ordinate in modo alfabetico.
ABEIMNOQSZ
ABDFJMNPQT
ABGKMNORUZ
ABDHLMNPRVW
Si supponga di utilizzare la tecnica dell'indicizzazione del prefisso, e di essere interessati ad una Jaccard Similarity **J=0,75** Si dica, giustificando le risposte:

- in cosa consiste la tecnica dell'indicizzazione del prefisso;
- quale deve essere la lunghezza del prefisso da considerare per l'indicizzazione delle varie stringhe;
- se essa può essere proficuamente applicata alle stringhe date;
- se esiste un criterio per scegliere un diverso ordinamento degli item che permette di migliorare tale tecnica. Se sì, qual è? In questo caso tale criterio porta beneficio?

Esercizio 3 (10 punti)

Si consideri l'algoritmo *CURE* per il clustering di punti in uno spazio euclideo con d dimensioni.
Si dica in modo **sintetico e puntuale,** giustificando le risposte:

a. Quali sono le assunzioni che tale algoritmo fa circa la conformazione dei cluster.
b. Come viene identificato e memorizzato un cluster
c. Se tale algoritmo è di tipo gerarchico e/o di tipo con assegnamento di punti

Infine, di descriva l'algoritmo evidenziando le sue varie fasi.

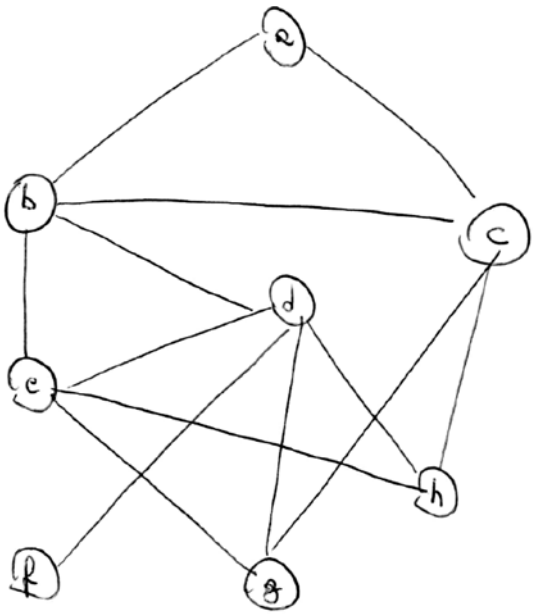
SECONDA PARTE

Esercizio 4 (17 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.



Esercizio 5 (17 punti)

Si consideri la stessa rete sociale dell'esercizio precedente.

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 tale calcolo può portare a dividere il grafo in due cluster, giustificando la risposta.

Regole per lo svolgimento della prova scritta:

- Per svolgere il compito si hanno a disposizione **40** minuti per la seconda parte e **110** minuti per l'intero compito
- 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.