

Cognome e Nome _____

Matricola _____

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

Appello del 21 giugno 2023

Esercizio 1 [8 punti]

Si progettino (eventualmente utilizzando più step di computazione tra loro concatenati) le funzioni *Map* e *Reduce* per restituire in output uno tutti i cicli di lunghezza 3 archi (quindi con 3 nodi), in un **grafo diretto** $G=(V, E)$. Si assuma che un arco del grafo di input, che va dal nodo i al nodo j , sia codificato con 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 [9 punti]

Si considerino gli insiemi descritti nella tabella qui sotto.

- **[4 punti]** Si scriva la colonna dell'insieme S_2 in modo che la Jaccard Similarity tra S_1 e S_2 sia $4/5$ e la Jaccard Similarity tra S_2 e S_3 sia $1/2$.
- **[5 punti]** Si calcoli, applicando il metodo visto a lezione ed evidenziando tutti i passaggi fatti riga per riga, la matrice delle firme minhash degli insiemi S_1, S_2, S_3 e S_4 secondo le permutazioni indotte dalle 3 funzioni hash evidenziate in tabella.

Riga	S_1	S_2	S_3	S_4	$(x+4) \bmod 7$	$(2x+3) \bmod 7$	$(3x-1) \bmod 7$
0	1		1	0	4		
1	1		0	0	5		
2	1		0	1	6		
3	0		0	1	0		
4	1		0	0	1		
5	0		0	1	2		
6	1		1	0	3		

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:

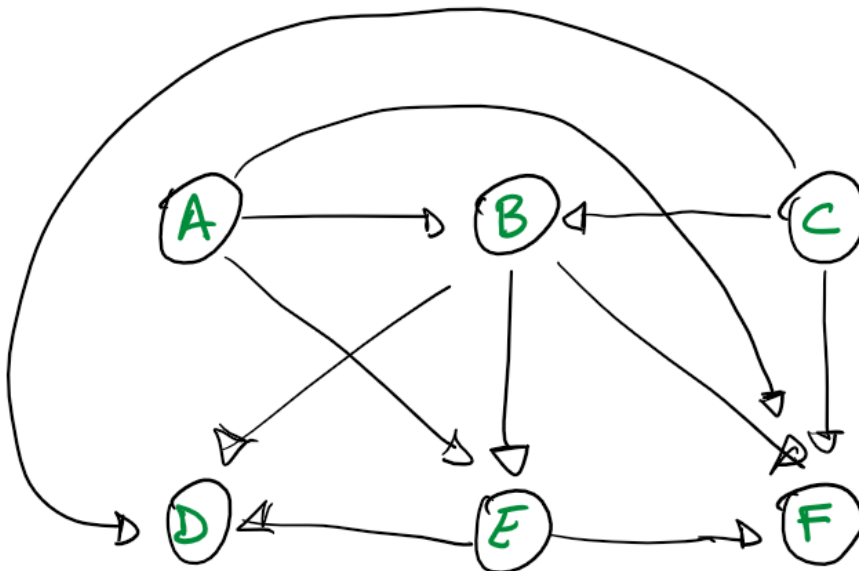
- Se tale algoritmo è di tipo gerarchico o di tipo con assegnamento di punti
- L'assunzione di fondo che questo algoritmo fa sui punti di ogni cluster
- Cosa sono i *discard set*, i *compressed set* e i *retained set* e perché vengono chiamati così.
- Come viene memorizzato in memoria un cluster, quanta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

Esercizio 4 [9 punti]

Si consideri la rete 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 **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.