Elective III BE (Comp) Sem VIII CBGS
Big Data Analytics.
VIII) CBGS 12/05/16

QP Code : 725400

(3 Hours) **[Total Marks :80**

N.B. : (1) Question no. **1** is **compulsory**
(2) Attempt any three from the remaining.
(3) Assume suitable data.

1. (a) What are the three Vs of Big Data? Give two examples of big data case studies. Indicate which Vs are satisfied by these case studies. **5**
   (b) Describe the operations of "shuffle" and "sort" in the Map reduce framework? Explain with the help of one example. **5**
   (c) Through an example illustrate how triples can be used to optimally store and count pairs in a frequent itemset mining algorithm. **5**
   (d) What is the motivation to count triangles in a social network graph? Outline one algorithm for counting triangles briefly. **5**

2. (a) What are the different data architecture patterns in NOSQL? Explain " Graph Store" and "Column Family Store" patterns with relevant examples. **10**
   (b) Show Map Reduce implementation for the following two tasks using pseudocode. **10**
   (i) Join of two relations with an example
   (ii) Multiplication of two matrices with one Map reduce step.

3. (a) Write a note on different distance measures that can be used to find similarity/ dissimilarity between data points in a big data set. **10**
   (b) Describe any two sampling techniques for big data with the help of examples. **10**

4. (a) Using an example bit stream explain the working of the DGIM algorithm to count number of 1's (Ones) in a data stream. **10**
   (b) Clearly explain how the CURE algorithm can be used to cluster big data sets. **10**

5. (a) Define Content based recommendation systems. Using an example case study describe how it can be used to provide recommendations to users. **10**

   (b) Let the adjacency matrix for a graph of four vertices (n_1 to n_4) be as follows: **10**

$$A = \begin{matrix} 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{matrix}$$

   Calculate the authority and hub scores for this graph using the HITS algorithm with k = 6, and identify the best authority and hub nodes.

[TURN OVER]

6. (a) Explain clearly how the SON partition based algorithm helps to perform frequent itemset mining for large datasets. How does this algorithm avoid False Negatives? **10**

   (b) For the graph given below use Clique percolation and find all communities **10**