35

12 May 2023

ISLAMIC UNIVERSITY OF TECHNOLOGY (IUT)

ORGANISATION OF ISLAMIC COOPERATION (OIC)

Department of Computer Science and Engineering (CSE)

SEMESTER FINAL EXAMINATION **SUMMER SEMESTER, 2021-2022**

DURATION: 3 HOURS **FULL MARKS: 150**

CSE 6279: Big Data Analysis and Management

Programmable calculators are not allowed. Do not write anything on the question paper.
Answer **all 6 (six)** questions. Figures in the right margin indicate full marks of the corresponding question.

---

1. a) What is meant by the "blind zone" in the context of data-mining? Rephrase the definition of Big Data by Gartner Inc. The definition highlights three major aspects. Briefly explain them. 5

   b) Consider one application was built using the traditional relational database system. The application addresses the following query: 10

   Find the list of people who are older than 35 years and do office at Gulshan area. Your tasks are as follows:

   - Outline briefly how would you design (in terms of ERD or DDLs) in a traditional database system driven application.
   - Now modify or add new features so that big data platform suits here. Also mention the major challenges in your new design.

   c) Explain base line of the Bonferroni's Principle (BP) to avoid "bogus" false positive. Consider the following scenario: 10

   Objective: To detect "evil doers", we hold the following assumptions:

   - There are 100 million people who might be evil doers.
   - Everyone goes to a hotel one day out of 200 days.
   - A hotel's capacity is 150 persons.
   - Total observation period is 300 days.
   - As a pattern for an evil doer we consider: "for a given hotel, 2 persons visit the hotel on 2 different dates for a common purpose".

   Your task is to apply the BP to test if this approach to detect evil doers is feasible.

2. a) What is shingle? State the problem of using simple Shingles for identifying Similar News Articles on web. Suggest an alternative to eliminate the problem. 10

   b) Briefly state the motivation of using Minhash Signature for document similarity measurement. 15

   Consider the Boolean matrix for 4 (d1,d2,d3,d4) documents as given in Table 1.

   Use **any four random permutations** to create the Minhash Signature. Show each step for the signature formation. Finally verify that the Jaccard Similarity of any 2 documents are very close to their Signature Similarity.

3. a) Can we construct Minhash Signature exploiting parallel processing? Justify your option. Briefly discuss the major problem of Minhash Signature for real implementation. 1

Table 1: Boolean Matrix for Q 2.b)

d1	d2	d3	d4
0	1	1	1
0	1	0	1
1	0	0	1
0	1	0	1
0	0	1	1
1	1	0	0

b) Consider the Boolean Matrix (against 4 Elements/Shingles) as shown in Table 2. Show each step to generate Minhash Signature based on the following two Random Hash Functions:

- $h1 = (x + 1)\%5$
- $h2 = (4x + 1)\%5$

Finally verify its correctness against Jaccard Similarity.

Table 2: Boolean Matrix for Q. 3.b)

Elements	d1	d2	d3	d4
a	0	1	0	1
b	0	1	0	0
c	1	1	0	1
d	0	1	1	1

4. a) Consider the following term-frequency vectors for 2 documents:

Table 3: Boolean Matrix for Q. 3.b)

	Computer	Network	Disk	RAM	CPU	BIOS	GPU
Document 1	0	1	0	0	0	0	0
Document 2	1	0	0	0	0	0	0

Apply Euclidean Distance to measure the similarity between these 2 documents. Is there any problem in this approach? Explain. Show the alternative solution to address the problem you mentioned.

b) Explain why Gaussian Elimination is not a feasible approach for the solution flow equation considering today's web structure.

c) Explain the problem statement of the flow equation in terms of Eigen Value and Eigen Vector. Present a motivating example of a network consisting of 4 nodes to explain the concept of flow equation formulation.

5. a) Briefly describe Spider-traps and Dead-ends events in page rank algorithm. Suggest suitable method to eliminate the problems.

b) *"In page rank algorithm, Spider-trap is not a problem while dead-end is a problem"* - Justify the statement.

c) Why do we need a different encoding scheme for storing the connectivity matrix M in page rank algorithm? Explain it using a network consisting of 5 nodes (with any connectivity you prefer) among them.

d) Present a comprehensive analysis of power iteration method for the following cases:

i. r^{new} fits in main memory and matrix M fits in disk

ii. r^{new} does not fit in main memory and matrix M fit in main memory

6. a) Describe the essential characteristics of a social network. Explain the Girvan-Newman Algorithm using suitable social graph example. 10

b) A social graph can be directed or undirected. Place two real-life examples in this regard. 15
Consider the following graph:

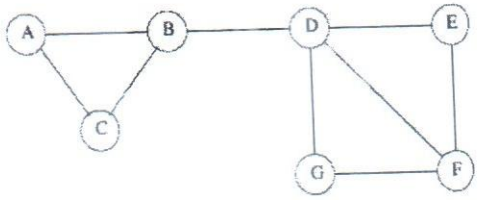

Figure 1: Graph for Question 4.b

The entities are the nodes A through G . The relationship, which we might think of as "friends," is represented by the edges. For instance, B is friends with A , C , and D .

Is this graph really typical of a social network, in the sense that it exhibits locality of relationships? Justify your position using step by step analysis.