THE UNIVERSITY OF AUCKLAND

SEMESTER TWO 2020

Campus: City, Offshore Online, UoA CLC - Northeast Forestry, UoA CLC - Southwest University

Algorithms for Massive Data

(Time Allowed: TWO hours)

NOTE:

- Attempt ALL questions in this exam.
- There are 50 marks in this exam.
- The exam counts 50% towards your final mark.

By completing this assessment, I agree to the following declaration:

I understand that the University expects all students to complete coursework with integrity and honesty. I promise to complete this online assessment with the same academic integrity standards and values. Any identified form of poor academic practice or academic misconduct will be followed up and may result in disciplinary action.

As a member of the Universitys student body, I will complete this assessment in a fair, hon
that:

- I declare tha
- I will not seek out any unauthorized help in completing this assessment.

- I declare that this work has not been submitted for academic credit in another University of Auckland course, or elsewhere.

- I am aware that the University of Auckland may use Turnitin or any other plagiarism detecting methods to check my content.
- I wi
an n
so

- I will not reproduce the content of this assess
form

1 Locality-sensitive Hashing [10 marks]

1.1 Computing MinHash signatures [5 marks]

Given 4 sets:

$$S_1 = \{3, 4, 5\}, S_2 = \{0, 1, 2\}, S_3 = \{0, 5\},$$
$$Q = \{0, 1, 2, 3, 4, 5\}.$$

1. Present these sets as a binary matrix where the set elements are $\{0, 1, 2, 3, 4, 5\}$. [**1 mark**]
2. Construct the MinHash signature matrix using 4 universal hash functions below. [**1 mark**]

   h_1 ,

   $h_3(x) = (x + 3) \texttt{ mod } 6,$ $\qquad h_4(x) = (x + 5) \texttt{ mod } 6$

3. Consider the set Q as the query set, estimate the Jaccard similarities $J(S_1, Q)$, $J(S_2, Q)$, and $J(S_3, Q)$. [**1 mark**]

4. Can we use the hash function in the form of $h(x) = (x + a) \texttt{ mod } 6$, where a is an integer to simulate random permutations for our sets? Explain you a [**2 marks**]

1.2 Tunin [2 marks]

Given the number of bands b and the num r, let $p = 1 - (1 - s^r)^b$ be the probability of being a candid

Jaccard similarity s.

Given the following values of r and b: $r = 3$ and $b = 10$; $r = 6$ and $b = 20$; $r = 5$ and $b = 50$, we compute the value p for s in range $\{0.1, 0.2, \ldots, 1\}$ as follows:

s	$(3, 10)$	$(6, 20)$	$(5, 50)$
0.1	0.0100	0.0000	0.0005
0.2	0.0772	0.0013	0.0159
0.3	0.2394	0.0145	0.1145
0.4	0.4839	0.0788	0.4023
0.5	0.7369	0.2702	0.7956
0.6	0.9123	0.6154	0.9825
0.7	0.9850	0.9182	0.9999
0.8	0.9992	0.9977	1.0000
0.9	1.0000	1.0000	1.0000

We would like to solve the **near neighbor search** problem using the Jaccard similarity. In particular, given a query set Q, we want to find **all** sets S_i such that $J(S_i, Q) \geq 0.5$. Which settings of b and r above should we use such that:

1. The probability that any 50%-similar pair is a candidate pair is at least 70%. Explain your solution. [**1 mark**]
2. The probability that any 50%-similar pair is a candidate pair is at least 70% and the number of candidate pairs is minimized. [**1 mark**]

1.3 Linear time of LSH on finding all similar pairs [3 marks]

Assume that the av
using the shingl
finding all Jaccard cu-
ments. In the lecture note, we state that "With LSH, we can approximately find all similar pairs in $O(n)$ time." Is the statement true or false? Explain your answer.

2 Streaming Algorithms [15 marks]

2.1 Reserv [5 marks]

Fig. 1.: Illustration of how pRS_1 works.

In our lecture, we have studied the reservoir sampling which samples an element from a stream of size m with the same probability. If we use the reservoir sampling with the summary size $s = 1$, each element of a stream will be sampled with probability $1/m$. We name this method as RS_1. The generalized version of reservoir sampling with the summary size $s > 1$ guarantees that each element in a stream will be sampled with the same probability s/m. We name this method as RS_s.

In the exam, we consider a new algorithm, called pRS_1, that simulates RS_s for $s > 1$ by running s independent RS_1 instances in parallel. pRS_1 also uses a summary of size s, as shown in Figure 1.

1. As a function of m and s, what is the probability an element of a stream is sampled by pRS_1? **[2 marks]**
2. Let f_i be the n ... plain how we can use ... **[3 marks]**

2.2 Misra-Gries vs. Reservoir Sampling **[5 marks]**

Run the Misra-Gries summary with $k = 2$ counters for the stream below:

$$\{1, 4, 5, 4, 4, 5, 4, 4\}$$

1. Present the final summary, including the elements and their counter values, whe ... **[1 mark]**
2. If we us ... $s = 2$ on this strea ... RS_s summary? **[2 marks]**
3. On Assignment 2, there is a request to "report th ... times the reservoir summary has been update ... of the summary size s and the stream le ... cted number of times the reservoir summary has been updated after processing the stream? **[2 marks]**

2.3 CountMin Sketch **[5 marks]**

Apply CountMin Sketch to estimate the frequency of each element in the stream below:

$$\{1, 4, 5, 4, 4, 5, 4, 4, 1, 4, 5, 4, 4, 5, 4, 4, 1, 4, 5, 4, 4, 5, 4, 4\}$$

Our CountMin Sketch uses $d = 3$ arrays with the hash functions as follows:

$$h_1(x) = (x + 1) \bmod 3,$$
$$h_2(x) = (3x + 1) \bmod 3,$$
$$h_3(x) = (5x + 2) \bmod 3.$$

1. Present the CountMin Sketch summary after processing all elements and the estimated frequency of each element. [**2 marks**]
2. Among Reservoir Sampling, Misra-Gries and CountMin Sketch, which algorithm we should use to find the top-1 frequent element in this stream. Explain your choice. [**3 marks**]

3 Algorithms for Large Graphs [15 marks]

3.1 Computing PageRank [5 marks]

Given the followin

$$\mathbf{A} = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix}$$

1. Convert $\mathbf{A}$ into a column-stochastic matrix $\mathbf{M}$. [**1 mark**]
2. In the lecture, we have shown that the PageRank $\mathbf{r} = \mathbf{M} \cdot \mathbf{r}$ is the eigenvector of the column-stochastic $\mathbf{M}$ corresponding to eigenvalue $\lambda = 1$. Com equa [**2 marks**]
3. Fro f running the power iteration algorithm $\mathbf{r} = \mathbf{M} \ \mathbf{r}$ on $\mathbf{M}$. Describe how to solve the problem. [**2 marks**]

3.2 Girvan-Newman [5 marks]

1. Compute the edge betweenness for all edges in the social network in Figure 2. Which edge will be removed to partition the graph into two parts using the Girvan-Newman method? [**3 marks**]

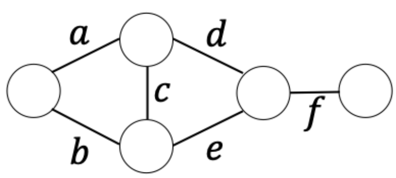

Fig. 2.: An example social network

2. In our lecture, we mentioned that we can use the Brandes' algorithm to calculate the shortest path from a node to all others. Does the algorithm apply for a weighted graph? Explain your answer. [**2 marks**]

3.3 Influence Maximization [5 marks]

1. Compute the influence spread of the seed set $S = \{a\}$ using the independent cascade model on the graph in Figure 3. **Hint:** Convert the stochastic graph to deterministic graphs. [**3 marks**]

Fig. 3.: A social network with activation probabilities on edges.

2. In the lecture, we gave the definition of submodular function as $f(S \cup \{v\})$ ... is the set of all item ... $f(A) + f(B) \geq$ $f(A$ ... o definitions are eq ... [**2 marks**]

4 Recommender Systems [10 marks]

4.1 Collaborative Filtering [6 marks]

Given the following transactions in the form of (user, item, rating) tuples in a recommender system.

$(u1, p1, 1.5)$, $(u1, p3, 4)$, $(u1, p5, 0.5)$, $(u2, p2, 4)$, $(u2, p4, 2)$, $(u3, p1, 4.5)$, $(u3, p4, 2.5)$, $(u3, p5, 5)$, $(u4, p2, 2)$, $(u4, p3, 3.5)$, $(u4, p4, 4)$, $(u4, p5, 2.5)$

Let the set of users be $\{u1, u2, u3, u4\}$ and the set of items be $\{p1, p2, p3, p4, p5\}$.

1. Construct the user-item interaction matrix based on the above transactions. Use question marks to denote missing values. [**1 mark**]
2. Apply the basic user-based collaborative filtering with the Pearson correlation coefficient for user $u2$ without considering bias. Give the top-1 recommended item to $u2$ and the corresponding predicted rating. [**2 marks**]

3. In the lecture, we discussed how to model the rating bias including the bias over all transactions, the bias of a user and the bias of an item. Give the predicted rating of user $u2$ to item $p5$ using the collaborative filtering that incorporates the above bias information. [**3 marks**]

Note: The predicted ratings should round to one decimal place.

4.2 Factorization Machine [4 marks]

Suppose you are asked to build a system to recommend events. Users $\{u1, u2, u3, u4\}$ attend events fro [illegible]
stadiums $s1$ and [illegible]

Table1.: Transactions of the event recommendation system

Transaction ID	Group of users	Event	Stadium
1	$u1, u2$	$e1$	$s1$
2	$u1, u3, u4$	$e2$	$s1$
3	$u2, u4$	$e3$	$s2$
4	$u3, u4$	$e1$	$s2$

1. Con [illegible] ng the event transactions in Table 1. [**1 mark**]
2. Can factorization machine predict the ratin[illegible] $u2$ may put on $e2$ held in $s2$? Explain your [illegible] [**1 mark**]
3. If we ignore the stadium information in the transactions and only consider users and events in the above example, does the factorization machine reduce to the latent factor model? If yes, explain your answer. If no, explain in what situation the factorization machine reduces to the latent factor model. [**2 marks**]