Enabling Large Scale Analytics: From Theory to Practice

June 17, 2016.

Name: ____
Instructions:

- Answer the questions directly on the exam pages.
- Show all your work for each question. Providing more detail including comments and explanations can help with assignment of partial credit.
- If you need extra space, use the back of a page.
- You can use the course notes.
- If you have questions during the exam, raise your hand.

Question	Value	Points Earned
1	10	
2	20	
3	10	
Total	40	

Question 1. (*10 points*) We have learnt algorithms for the following problems in the class.

1. Bloom Filter
2. Count-Min sketch
3. Min-wise independent hashing
4. Locality sensitive hashing
5. Dense subgraph detection

Indicate which of the above are applicable in the following scenarios. No justification is required.

1.1 (*2 points*): *A journal editor wants to check quickly for plagiarism for every newly submitted article.*

Min-wise independent hashing/ Locality sensitive hashing

1.2 (*2 points*): *In computing, a denial-of-service (DoS) attack is an attempt to make a machine or network resource unavailable to its intended users, such as to temporarily or indefinitely interrupt or suspend services of a host connected to the Internet. This is done by sending a large volume of packets to the victim destination from multiple spoofed host ids.*

Count-Min Sketch

1.3 (*2 points*): *When sending an email to a client, you want the mail server to quickly check if the email address has been used previously.*

Bloom Filter

1.4 (*2 points*): *In a protein-protein interaction network, there is an edge between every pair of proteins. A highly connected subgraph in this network often refers to a protein complexes which need to be detected.*

Dense Subgraph Detection

1.5 (*2 points*): *The entire human genome which is a sequence of* 4 *unique characters* A, C, G, *and* T *is broken into smaller chunks and stored in a data structure to rapidly find segments which may contain high similarity to a known genetic mutation. The known genetic mutations can be provided any time as a sequence of* A, C, G, *and* T *characters.*

Locality sensitive hashing/ Min-wise independent hashing

Question 2. (*20 points*) In this question, you will have to show the output of various algorithms that you have learnt in the course.

2.1 (*6 points*): *Show the execution of Count-Min sketch data structure on the following input. Draw the Count-Min sketch table.*

$$2, 2, 15, 1, 10, 1, 1, 2, 15, 2$$

Assume there are 13 cells in each hash table. Use the following two hash functions:

1. $h_1(x) = (5 + 9x)\ mod\ 29\ mod\ 13$
2. $h_2(x) = (4 + 7x)\ mod\ 29\ mod\ 13$

0	3	0	0	0	0	0	0	1	0	4	2	0
0	0	0	1	0	4	0	0	0	2	0	3	0

2.2 (*2 points*): *What are the estimated frequencies for* 2*,* 15*,* 1 *and* 10*?*

$4, 2, 3$ and 1

2.3 (*6 points*): *Compute the Jaccard similarities of each pair of the following three sets:* $\{1, 2, 3, 4\}$*,* $\{2, 3, 5, 7\}$*, and* $\{2, 4, 6\}$*. Consider the following permutation* $10, 3, 1, 4, 2, 6, 8, 5, 9, 7$*, and compute the minhashes of the three sets based on this permutation.*

$\{1, 2, 3, 4\}$, $\{2, 3, 5, 7\}$: $\frac{2}{6} = \frac{1}{3}$
$\{2, 3, 5, 7\}$, $\{2, 4, 6\}$: $\frac{1}{6}$
$\{1, 2, 3, 4\}$, $\{2, 4, 6\}$: $\frac{2}{5}$

Min-hashes: $3, 3, 4$

2.4 (*2 points*): *Consider the following locality sensitive hashing for* 10*-dimensional binary vectors,* $\mathcal{H} = \{h_1, h_2, h_3, h_4, .., h_{10}\}$*, where* h_i *returns the ith bit of the vector,* $i = 1, 2, 3, 4, .., 10$*. Compute* $Prob_{h\sim\mathcal{H}}(h(x) = h(y))$ *when* $x = 0110111100$ *and* $y = 1100111000$*, and also when* $x = 0110111100$ *and* $y = 0001000101$*.*

$$Prob_{h\sim\mathcal{H}}(h(x) = h(y) \mid x = 0110111100, y = 1100111000) = \frac{7}{10}$$

$$Prob_{h\sim\mathcal{H}}(h(x) = h(y) \mid x = 0110111100, y = 0001000101) = \frac{3}{10}$$

2.5 (*4 points*): *What is the expected density of a random graph where there is an edge between any pair of vertices with probability* p*? Express the expected density in terms of number of vertices of the graph and* p*.*

$\frac{(n-1)p}{2}$ where n is the number of vertices.

Question 3. (*10 points*) In this question, you will be tested on simple probability concepts.

Suppose X is the number of dust storms that occur on Mars next year. You should assume that X is a discrete uniform random variable that take one of the 101 values in the range $\{0, 1, 2, 3, \ldots, 100\}$. Let $Y = |X - E(X)|$.

3.1 (*3 points*): *Enter values for the following probabilities:*

$$E(X) = \frac{1}{101}\frac{100 * 101}{2} = 50$$

$$var(X) = \frac{1}{101}\frac{100 * 101 * 201}{6} - 50^2 = 850$$

$$P(X = 12) = 1/101$$

3.2 (*2 points*): *Enter the following values. You may use the fact* $1 + 2 + \ldots + 50 = 1275$.

$$P(Y = 0) = 1/101$$

$$P(Y = 1) = 2/101$$

$$P(Y = 2) = 2/101$$

$$E(Y) = (1 + 2 + \ldots + 50) \times 2/101 = 2550/101$$

3.3 (*2 points*): *By applying the Markov Bound to* Y*, give an upper bound for the following quantity:*

$$P(|X - E(X)| \geq 30) = P(Y \geq 30) \leq E(Y)/30 = 2550/3030 = 85/101 = 0.84\ldots$$

3.4 (*2 points*): *By applying the Chebyshev Bound, give an upper bound for the following quantity:*

$$P(|X - E(X)| \geq 30) \leq var(X)/30^2 = 850/900 = 17/18 = 0.94\ldots$$

3.5 (*1 points*): *What is the exact value of* $P(|X - E(X)| \geq 30)$*?*

$$P(|X - E(X)| \geq 30) = P(Y = 30) + P(Y = 31) + \ldots + P(Y = 50) = 21 \times 2/101 = 42/101 = 0.41\ldots$$