Foundations of Data Science

For Ph.D. Qualifying Exam at DaSE (2020)

Name: ____________________ **Student ID:** _______________ **Credits:** ________________

Read all of the following information before starting the exam:

- Show all work, clearly and in order, if you want to get full credit. I reserve the right to take off points if I cannot see how you arrived at your answer (even if your final answer is correct).
- Justify your answers whenever possible to ensure full credit. When you do use your calculator, sketch all relevant graphs and explain all relevant mathematics.
- Circle or otherwise indicate your final answers.
- Please keep your written answers brief; be clear and to the point. I will take points off for rambling and for incorrect or irrelevant statements.
- This test has 8 problems and is worth 100 points, plus some extra credit at the end. It is your responsibility to make sure that you have all of the pages!
- Good luck!

1. (*10 points*) Let following matrix A be the adjacency matrix of graph G. Please write down the transition probability of the random walk, Laplacian matrix and normalized Laplacian matrix corresponding to G.

$$\begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{pmatrix}$$

2. (*10 points*) In n tosses of a fair coin, let X be # heads, what is the upper bound of $P(X > \frac{5n}{6})$ given by following inequalities?

a. Markovs inequality;

b. Chebyshevs inequality;

c. Chernoff bound.

3. (*10 points*) As shown in the following table, given a universal set U of five elements, there are three subsets S_1, S_2 and S_3.

a. Compute the Jaccard similarity of each pair of columns.

b. Compute the minhash signature for each column if we use the following hash functions: $h_1(x) = 7x + 1 \bmod 6$; $h_2(x) = 11x + 3 \bmod 6$; $h_3(x) = 5x + 2 \bmod 6$, where $x \in U$.

c. Compute similarity of each pair of sets via using the minhash signatures.

Element	S_1	S_2	S_3
0	1	1	1
1	0	0	0
2	1	1	0
3	0	1	1
4	1	0	1
5	0	0	0

4. (*10 points*) Let A be the adjacency matrix of graph G, where

$$A = \begin{pmatrix} 0 & 4 & 0 & 0 \\ 4 & 0 & 1 & 0 \\ 0 & 1 & 0 & 4 \\ 0 & 0 & 4 & 0 \end{pmatrix}$$

Once we have community structure, the modularity of the community can be computed as

$$Q = \frac{1}{2m} \sum_{i,j} (A_{ij} - \frac{k_i k_j}{2m}) \delta(C_i, C_j),$$

where m and C_i denote # edges and the $i-$th community in the graph, k_i is the degree of vertex v_i, and

$$\delta(C_i, C_j) = \begin{cases} 1, & \text{if } C_i = C_j; \\ 0, & \text{otherwise.} \end{cases}$$

a. If there are two communities: $C_1 = \{1, 2\}$ and $C_2 = \{3, 4\}$, please compute the modularity of the partition;

b. Is there any way to increase the value of modularity via adjusting the community structure?

5. (*10 points*) Given a universal set $U = \{a, b, c, d, e, f, g, h, i, j, k, l\}$, the set $S = \{A_1, \cdots, A_7\}$ contains the following subsets of U, i.e., $A_i \subset U$.

$$A_1 = \{a, b, c, d\}, A_2 = \{e, f, g, h\}, A_3 = \{i, j, k, l\}$$
$$A_4 = \{a, e\}, A_5 = \{i, b, f, g\}, A_6 = \{c, d, g, h, k, l\}, A_7 = \{l\}$$

a. Please find a cover of set U;

b. Using Hill-Climbing algorithm to find minimal cover of universal set U.

6. (*10 points*) A certain experiment is believed to be described by a two-state Markov chain with the transition matrix P, where $P = \begin{pmatrix} 0.5 & 0.5 \\ p & 1-p \end{pmatrix}$ and the parameter p is unknown. When the experiment is performed many times, the chain ends in state one approximately 20 percent of the time and in state two approximately 80 percent of the time.

a. Compute a sensible estimate for the unknown parameter p and explain how you found it;

b. Whether is the Markov chain irreducible, or not? Why?

c. Whether is the Markov chain aperiodic, or not? Why?

7. (*20 points*) Given a set V and $A \subseteq V$, let $f(A)$ be a set function. If $f(A)$ is a submodular

a. What conditions does $f(A)$ satisfy?

b. Fixing set $S \subset V$, prove that $g(A) = f(A \cap S)$ is a submodular.

c. Fixing set $S \subset V$, prove that $h(A) = f(A \cup S)$ is also a submodular.

8. (*20 points*) Given an input stream $< 4, 1, 3, 5, 1, 3, 2, 6, 7, 1, 8, 1 >$ and hash functions in the form of $h(x) = (ax + b) \bmod 4$, where a and b are two arbitrary integers. If there are three following hash functions:

(1) $h(x) = (3x + 2) \bmod 4$;

(2) $h(x) = (7x + 5) \bmod 4$;

(3) $h(x) = (5x + 3) \bmod 4$;

Please address the following questions:

a. Find the frequency count of every item given by Count-Min sketch;

b. According to the example, analyze the pros and cons of Count-Min sketch;

c. Analyze the accurate of counting result in a.;

d. If we try to find the (ϵ, δ)-approximations of the frequencies, how to modify the algorithm;