Quiz 5: Similar Items

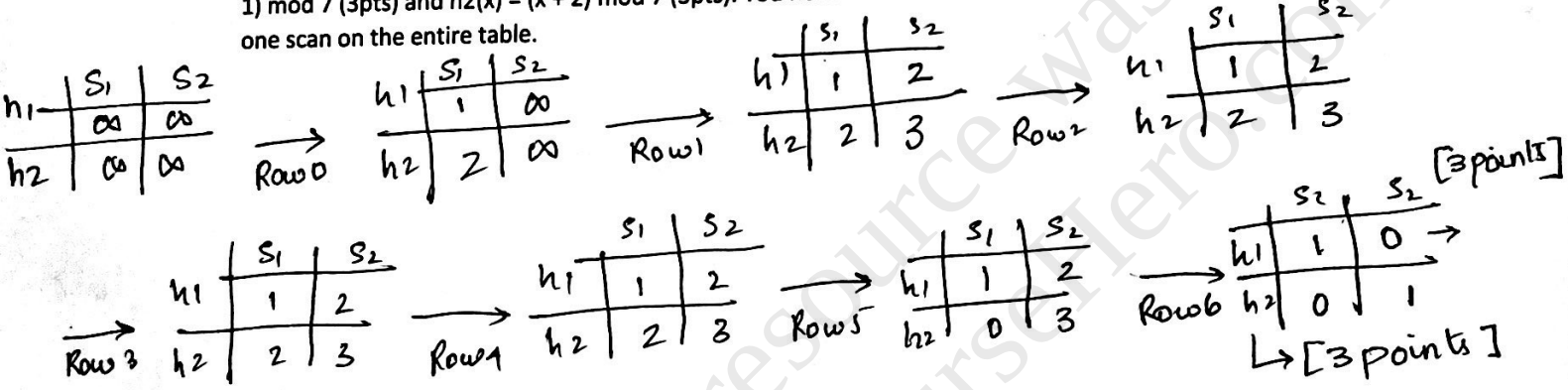
Name: GRADER ID: _____

Consider the following characteristic matrix of two sets: S1 and S2.

Row #	S1	S2
0	1	0
1	1	1
2	1	1
3	0	0
4	1	1
5	1	0
6	0	1

H1	H2
1	2
2	3
3	4
4	5
5	6
6	0
0	1

1. [6 points] What are the two minhash values of S1 and S2 based on the permutation using h1(x) = (x + 1) mod 7 (3pts) and h2(x) = (x + 2) mod 7 (3pts). You need to calculate the minhash values with only one scan on the entire table.



2. [4 points] Construct a signature for S1 and S2 based on the minhash values obtained from h1(x) and h2(x) above. Estimate the Jaccard similarity of S1 and S2 using the signature. What is the actual Jaccard similarity of S1 and S2 (2pt)? Is the estimate close to the actual Jaccard similarity? If not, suggest a way to improve the estimate (2pts).

Actual Jaccard similarity of S_1 and S_2 = 3/6 = 1/2

[1 point]

Jaccard similarity using signature = 0/2

[Estimated] [1 point]

NO, the estimate is not close to the actual Jaccard similarity. To improve, increasing the number of permutations or hash functions will help

[2 points]