Name: _____ USC ID: _____

Quiz 5: Finding Similar Sets (10 points), 15 minutes

TA handling this quiz: Yiwen Tang yiwentan@usc.edu

Consider three sets: S1, S2, and S3 whose characteristic matrix is shown in columns 1-4 of the table below.

Row #	S1	S2	S3	f1(x)=(x+1)%5	f2(x)=(3x+1)%5
0	0	1	0	1	1
1	1	1	0	2	4
2	0	0	1	3	2
3	1	1	1	4	0
4	0	0	1	0	3

1. [2 points] Suppose we generate signatures of the sets using hash functions f1(x) and f2(x) shown in the table above. Recall that x is the old row #. Fill in the blanks in the above table with the new row numbers generated by each hash function.

2. [5 points] Consider building a signature matrix using the one-pass (through the characteristic matrix) algorithm discussed in class. Fill in the blanks below with the updated content of signature matrix after processing (old) row #0, #1, and so on.

	S1	S2	S3
f1	∞	∞	∞
f2	∞	∞	∞

	S1	S2	S3
f1	∞	1	∞
f2	∞	1	∞

(0)

	S1	S2	S3
f1	2	1	∞
f2	4	1	∞

(1)

	S1	S2	S3
f1	2	1	3
f2	4	1	2

(2)

	S1	S2	S3
f1	2	1	3
f2	0	0	0

(3)

	S1	S2	S3
f1	2	1	0
f2	0	0	0

(4)

3. [3 points] Use the signatures obtained above, estimate the **pairwise** similarity of the 3 sets. For each pair, indicate if the estimate is over- or under-estimate.

Pair	Actual Sim.	Estimated Sim.
(S1, S2)	2/3	1/2
(S1, S3)	1/4	1/2
(S2, S3)	1/5	1/2

Sim(S1, S2) is under-estimate, Sim(S1, S3) and Sim(S2, S3) is over-estimate.