Data Mining Midterm – Yaffa Atkins

1) Design map reduce to return the min

Map:

a. For each t in tuple
    i. Return (1,t)
    ii. t represents a list [t1 t2 t3… tN] of all ints in a group

Reduce:

b. Mins = []
c. Min = infinity
d. For each group
    i. Return min([t1 t2 t3… tN])
    ii. Append min to mins

For int in min

If min[int] < min

Min = If min[int]

Return min

2)

1-(1-.81^10)^50 = probability of being a candidate pair 99.8%

1-(1-.69^10)^50 = probability of being a candidate pair 71%

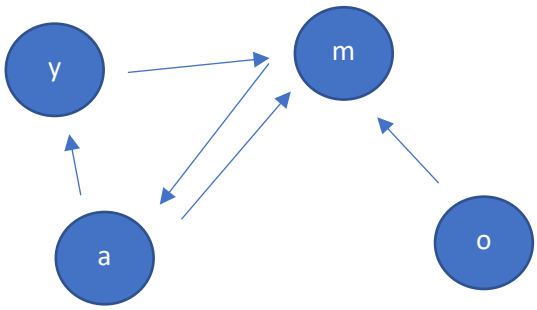
3) (fraction of ones)^ # of hash functions

Say there are 1000 bits, 5 hash functions, inserting 20 bits

T = 500, d = 5*20

Probability of there being a false positive = ((5*20)/500)^5 = 0.00032 = 3.2%

4)



Y = a/2 because a points to and one other place

A = m

M = a/2 + y + 0

O = 0 because nothing points to it

If y + a + m + o = 1, then y = ½, a = 1/3, and m = 1/9 and o = 0.

5)

Sets:

S1 = {1,0,0,0,1,0,0}

S2 = {0,1,1,0,0,1,0}

S3 = {0,0,0,0,1,0,1}

S4 = {1,1,0,1,0,1,0}

Calculate Jaccard similarity of all the Pairs:

S1 and s2 = 2/7

S1 and s3 = 5/7

S1 and s4 = 3/7

S2 and s3 = 2/7

S2 and s4 = 4/7

S3 and s4 = 1/7

3 hash functions

h1: 1X + 3 mod 7

h2: 3x + 1 mod 7

h3: 5x + 2 mod 7

Sig matrix:

	S1	s2	s3	s4
H1	0	4	0	3
H2	1	0	2	1
H3	1	0	1	0

6) no