CS345A Final Examination

March 19, 2009, 12:15 - 3:15PM

Name: _____

I acknowledge and accept the Honor Code.

_____ (signature)

Directions: Answer all 14 questions on the exam paper itself. The total number of points is 180 (i.e., one point per minute). The exam is open book/notes. Computers or calculators may be used.

Question	**Max Pts.**	**Score**
1	10	
2	10	
3	15	
4	10	
5	15	
6	20	
7	10	
8	10	
9	15	
10	15	
11	10	
12	15	
13	15	
14	10	
Total	180	

Question 1: (10 pts.) Suppose we represent sets by ordered strings and index a prefix of the appropriate length. If the string *abcdefgh* indexed under *a*, *b*, and *c*, but not *d* through *h*, what is the range of possible values for the maximum Jaccard distance J?

Lower bound: _____

Upper bound: _____

Question 2: (10 pts.) Suppose we have a family **H** of (0.1, 0.9, 0.6, 0.4)-sensitive hash functions.

a) If we apply the 2-way AND construction (i.e., construct hash functions that say "yes" if and only if a pair of hash functions from **H** both say "yes"), what is the sensitivity of the resulting family? (That is, give the parameters a, b, c, and d such that the result is an (a,b,c,d)-sensitive family.)

_____

b) If we apply the 3-way OR construction to H (not to the result of part a), what is the sensitivity of the resulting family?

_____

Question 3: (15 pts.) The following is a matrix representing three sets, X, Y, and Z, and a universe of five elements *a* through *e*.

Row	X	Y	Z
a	0	0	1
b	1	1	1
c	0	1	1
d	1	0	0
e	0	1	0

a) Give the Jaccard similarities of each pair of sets:
sim(X,Y) = _____ sim(X,Z) = _____ sim(Y,Z) = _____
b) Suppose we create minhash signatures of length 5 for each of the three sets X, Y, and Z. The signatures are based on the five cyclic permutations of the rows. That is, the first permutation uses order abcde, the second uses bcdea, the third uses cdeab, the fourth deabc, and the fifth eabcd. Give the signature matrix below.

Perm.	X	Y	Z
1			
2			
3			
4			
5			

c) What are the estimated Jaccard similarities for each pair of sets according to the signatures of the sets?

est-sim(X,Y) = _____ est-sim(X,Z) = _____ est-sim(Y,Z) = _____

Question 4: (10 pts.) We wish to run the PCY algorithm on a data set with a billion baskets. Each basket contains n items. On the first pass, we can afford to store in main memory a billion integers, each of which is a bucket.

a) As a function of n, what is the maximum support threshold s we can allow if the average count for a bucket is to be no more than half the threshold?

_____

b) Suppose we use the multihash extension to PCY, in which we divide the available space for buckets among m hash tables and hash each pair to each table. Assuming $n = 10$, i.e., there are 10 items per basket, and assuming that we want the average bucket count in each hash table to be no more than half the support threshold s, what is the relationship between h and s?

_____

Question 5: (15 pts.) Suppose that A, B, C, D, E, and F are all the items. For a particular support threshold, the maximal frequent itemsets are {A,B,C} and {D,E}.

a) (5 pts.) What are all the other frequent itemsets?

_____

b) (10 pts.) What is the negative border?

_____

Question 6: (20 pts.) Suppose we apply the AMS version of the Flagolet-Martin algorithm described in class. We shall choose our hash functions from the family $h_i(x) = (x+i) \bmod 2^{32}$. The stream for which we want to estimate the number of different elements has only the elements 3, 5, and 8, repeated many times.

a) What is the estimate of the number of distinct elements if we use the hash function h_0?

_____

b) What is the estimate of the number of distinct elements if we use the hash function h_1?

_____

c) Give an example of a hash function h_i (just give the value of i) that gives the minimum possible estimate of the number of distinct elements.

_____

d) Briefly explain your answer to (c). Why do you get that estimate for your chosen value of i? Why is it the minimum possible?

_____

_____

_____

_____

Question 7: (10 pts.) Suppose we use the AMS algorithm described in class to estimate the second moment (surprise number) of the string *abacadbcab* and we construct three random variables X_1, X_2, and X_3, based on the (randomly chosen) positions 3, 5, and 8 of the string. What are the values of these variables?

X_1 = _____ X_2 = _____ X_3 = _____

Question 8: (10 pts.) Below is a table of distances between four points *a, b, c*, and *d* in a non-Euclidean space.

	b	*c*	*d*
a	3	10	4
b		7	8
c			6

Suppose {*a,b,c,d*} is a cluster, which we want to represent by its clustroid.

a) Give an example of a commonly used clustroid definition that makes *a* the clustroid.

_____

_____

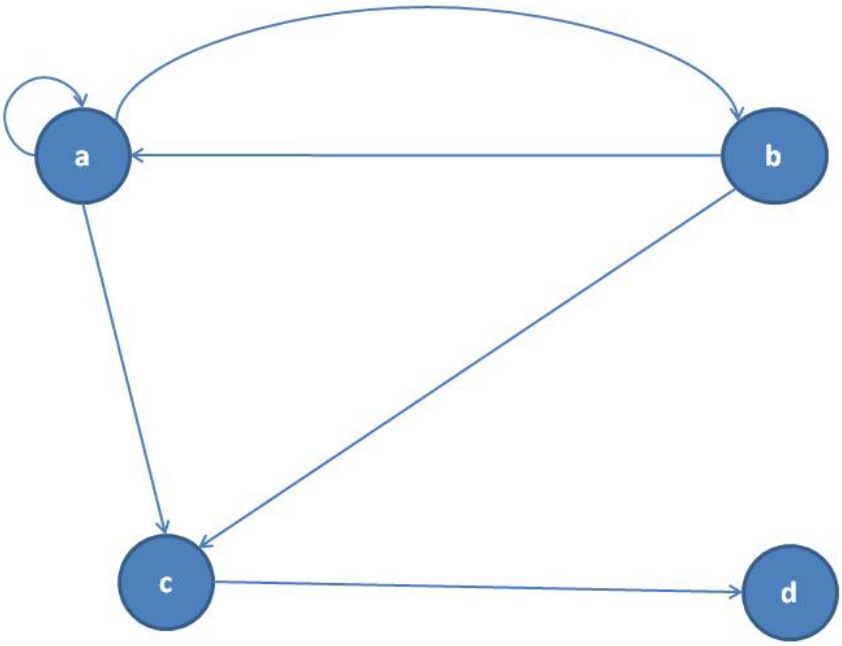
b) Give an example of a commonly used clustroid definition that makes *b* the clustroid.

_____

_____

Question 9:(15 pts.)

You are going to calculate the apply the basic PageRank algorithm to the below network:



a) Write down the final column-stochastic matrix used for PageRank score calculation, using β = 0.8 for random teleportation, and pruning and removing all dead ends (if any). (5 pts.)

b) Compute the PageRank scores of all below nodes, approximating values for deadends (if any) by propagating from a reduced graph (10 pts.):

a:_____ **b**: _____ **c**: _____ **d**: _____

Question 10 (15 points). Consider the "host size" example discussed in the Map Reduce lecture. The input file has records of the form <URL, size>, and you are given a function

url_host(URL) that, given a URL, returns the hostname. The goal is to compute the sum of the sizes of all the web pages for each hostname in the input file.

Suppose we use Map Reduce with 10 map workers and 5 reduce workers. Assume that the map reduce implementation schedules the map tasks on the nodes that contain chunks of the input file on their local disks, and that no two tasks (map/map, map/reduce, reduce/reduce) are scheduled on the same node. The distributed files system uses 2-way replication. You can assume that one replica of each chunk will be written on the same node as the task that produces it.

The input file contains 100 million records, representing 10 million unique hosts, and each chunk contains on average 3 million unique hostnames. The average length of a URL is 100 bytes, the average length of a hostname is 20 bytes, and sizes are encoded using 4-byte integers.

(a) Assume that we use a map-reduce implementation without a combiner. Estimate the total disk I/O and network I/O during the computation.

Disk I/O = _____

Network I/O = _____

(b) Suppose we use a combiner to reduce the network I/O. Estimate the total disk I/O and network I/O during the computation.

Disk I/O = _____

Network I/O = _____

Question 11 (10 points). Consider a Page Rank computation over a graph with 100 million web pages. Each page has on average 10 outlinks. The graph is encoded using the sparse-matrix encoding discussed in class. Each entry in the encoding occupies 4 bytes, as does each entry in the page-rank vector.

(a) What is the size of the graph on disk? _____

(b) If we use a computer with 1 GB of RAM, which of the computation methods discussed in class would you use?

_____

(c) Estimate the total I/O if the computation takes 50 iterations.

_____

Question 12 (15 points). Using the DIPRE/Snowball algorithm to mine tuples, we find the following patterns:

- Pattern p, with 1000 positive matches and 250 negative matches.

- Pattern q, with 750 positive matches and 300 negative matches.
- Pattern r, with 100 positive matches and 10 negative matches.

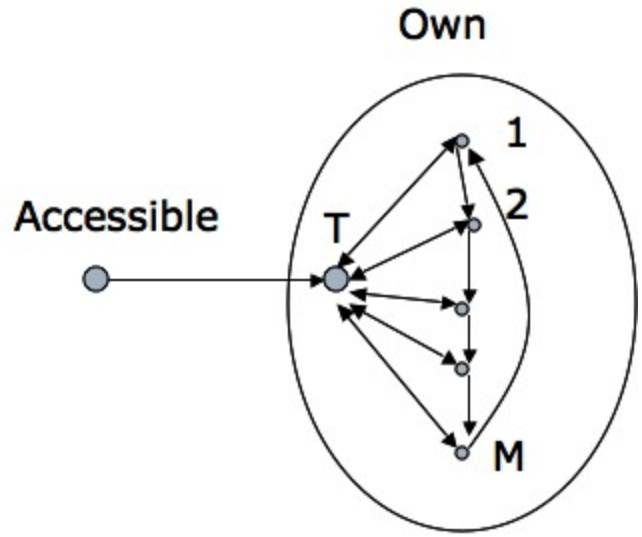
(a) Compute the confidence of each pattern.

conf(p) = _____ conf(q) = _____ conf(r) = _____

(b) Tuple t matches patterns p and q; tuple u matches pattern p alone; while tuple v matches all three patterns. Compute the confidence of each tuple.

conf(t) = _____ conf(u) = _____ conf(v) = _____

Question 13 (15 points).



A clever spammer uses a variant of the Link Farm topology discussed in class. The spammer gets a link to the target page T from an accessible page A with page rank a; the link to page T is one of 10 outlinks from A. The link farm contains M pages linked in a directed cycle as shown. The total number of pages on the web is N, and the teleport parameter is b.

Assume N = 1 billion, a = 10^{-7}, M = 100,000 and b = 0.85.

(a) What is the page rank of each "link farm" page? _____

(b) What is the page rank of the target page? _____

Question 14 (10 points). Consider a search engine advertiser auction involving 3 advertisers A, B and C. Generalized BALANCE is used to determine one advertiser per query.

Advertiser	**Bid**	**CTR**	**Budget**	**Spent so far**
A	$1	10%	$1000	$100
B	$2	8%	$2000	$1000
C	$3	4%	1000	$400

(a) If a query arrives that is bidded on by A and B, the winner is: _____

(b) If a query arrives that is bidded on by A and C, the winner is: _____

(c) If a query arrives that is bidded on by A, B, and C, the winner is: _____