

**CS 345A Data Mining
Final Exam – Autumn 2005**

This exam is open book and notes. You have 180 minutes to complete it.

Print your name: _____

The Honor Code is an undertaking of the students, individually and collectively:

1. that they will not give or receive aid in examinations; that they will not give or receive unpermitted aid in class work, in the preparation of reports, or in any other work that is to be used by the instructor as the basis of grading;
2. that they will do their share and take an active part in seeing to it that others as well as themselves uphold the spirit and letter of the Honor Code.

The faculty on its part manifests its confidence in the honor of its students by refraining from proctoring examinations and from taking unusual and unreasonable precautions to prevent the forms of dishonesty mentioned above. The faculty will also avoid, as far as practicable, academic procedures that create temptations to violate the Honor Code.

While the faculty alone has the right and obligation to set academic requirements, the students and faculty will work together to establish optimal conditions for honorable academic work.

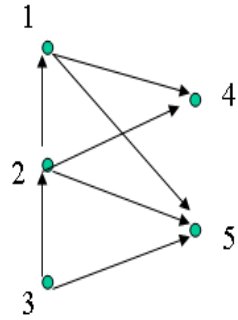
I acknowledge and accept the Honor Code.

Signed: _____

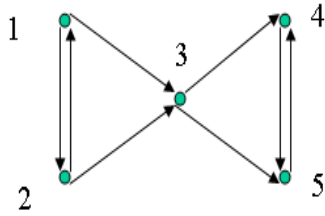
Problem	Points	Maximum
1		10
2		20
3		10
4		10
5		10
6		10
7		10
8		19
9		13
10		8
11		8
12		13
13		13
14		8
15		19
Total		180

Problem 1 (10 points)

Set up the hubs-and-authorities equations for the graph shown in the figure. Denote the hub score and authority score of node i by h_i and a_i , respectively.



Problem 2 (20 points)



(a) Set up the equations to compute pagerank for the graph shown in the figure. Assume that the “tax” rate (i.e., the probability of teleport) is 0.2.

(b) Set up the equations for topic-specific pagerank for the same graph, with teleport set 1,2. Solve the equations and compute the rank vector.

- (c) Suppose we use the inverse page rank approach to find a seed set of size 2 for TrustRank. Which nodes would we pick?

Problem 3 (10 points)

Suppose you are given the the following TSPR vectors computed on web graph G , but you are **not** allowed to access the graph itself.

- r_1 , with teleport set 1,2,3
- r_2 , with teleport set 3,4,5
- r_3 , with teleport set 1,4,5
- r_4 , with teleport set 1

Is it possible to compute each of the following rank vectors without access to the web graph G ? If so how? If not why not? Assume a fixed teleport parameter β .

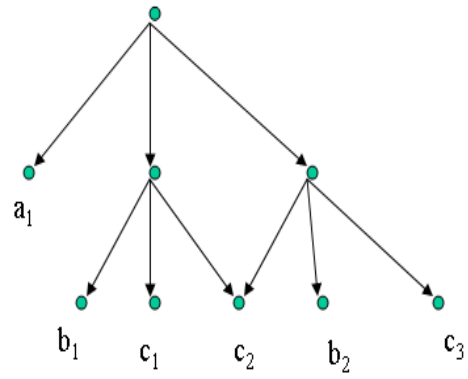
(a) r_5 , corresponding to the teleport set 2

(b) r_6 with teleport set 5

(c) r_7 , with teleport set 1,2,3,4,5, with weights 0.1,0.2,0.3,0.2,0.2 respectively.

Problem 4 (10 points)

Is there a Perfect Compact Skeleton (PCS) for the data graph shown in the figure? Assume the schema has three attributes, A, B, and C, and data values of the form x_i correspond to attribute X. If there is a PCS, what are the tuples in the relation for the data graph? If there is no PCS, explain why there is none.



Problem 5 (10 points)

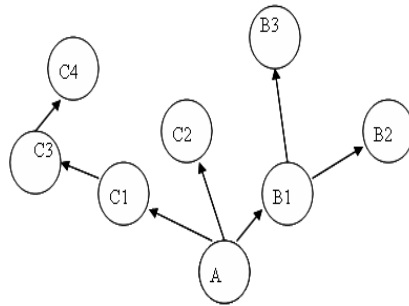
Consider the AdWords problem with three advertisers A_1 , A_2 and A_3 and three kinds of queries, X, Y and Z. Each advertiser has a budget of 3. * A_1 bids 1 on query X and 0 on queries Y and Z * A_2 bids 1 on X and Y and 0 on Z * A_3 bids 1 on X, Y and Z

(a) What is the competitive ratio of the Greedy algorithm for this configuration?

(b) What is the competitive ratio of the BALANCE algorithm for this configuration?

Problem 6 (10 points)

All parts in this question will use the following graph, which shows Web pages and their link structure. There are three domains in this graph: A, B, and C. A has 1 Web page. B has 3 Web pages: B1, B2, B3. C has 4 Web pages: C1, C2, C3, C4.



- (a) Assume that the number of changes of page B1 is modeled as a Poisson process parameterized by $\lambda = 3$ changes / hour. Suppose that a crawler visited B1 at 6:00AM, 6:30AM, and 6:35AM and found that the page's content at 6:30AM was different from that at 6:00AM. The crawler also found that the page's content at 6:35AM was the same as that at 6:30AM. When should the crawler crawl B1 again in order to detect its next change? Explain your answer in one or two sentences.

Consider the shown Web graph. Suppose that it takes a crawler 100 milliseconds to fetch one Web page. Suppose further that politeness policy states that, given any domain, a crawler cannot issue more than 1 page request to it in 100 milliseconds. This means that, to crawl two pages x and y in the same domain, 100 milliseconds must have elapsed between the end of crawl on page x and the beginning of crawl on page y .

- (b) Crawler X must crawl in the following manner: once the crawler fetches a page in a domain, the crawler must finish crawling all pages within that domain before switching to a different domain. What is the shortest time in which X can finish crawling all pages if it observes the politeness policy and starts crawling on page A? Give us your numerical answer and an explanation of how X can achieve your answer.

- (c) Crawler Y crawls in the following manner: it can crawl any Web page that is linked to by a Web page that has already been crawled. If Y also observes the politeness policy and starts crawling on page A, what is the shortest time in which Y can finish crawling all pages? Give us your numerical answer and the crawling order of pages that achieves this answer.

- (d) Suppose that all web pages change content once every 150 milliseconds. That is, they are simultaneously modified at 1:00AM, 1:00AM 150ms, 1:00AM 300ms ... (this is possible if these pages are dynamically generated). Given the crawling order that you propose in (c), suppose that the crawler starts crawling by sending request for page A at 1:00AM 1ms. What is the freshness and age of the pages already collected in the search engine database at 1:00AM 202ms? Show your calculations.

Problem 7 (10 points)

(a) Consider two documents A and B. Each document's number of tokens is $O(n)$. What is the runtime complexity of computing A and B's k-shingle resemblance? Assume that comparison of two k-shingles to assess their equivalence is $O(k)$. Express your answer in terms of n and k.

(b) Given two documents A and B, if their 3-shingle resemblance is 1, does that mean A and B are identical? If so, prove it. If not, show a counterexample.

(c) Is there an $n \geq 1$ such that the following statement is true?

Two documents A and B with n-shingle resemblance equal to 1 must be identical.

If so, provide the smallest such n and show why. If not, state how you can construct counterexamples for each n.

Problem 8 (19 points)

In this question assume:

- All integers and item ID's require 4 bytes.
- The machine being used has 240 million bytes of available main memory.
- There are 1 million items, of which 100,000 are frequent.
- There are 1 million pairs of items that are frequent pairs.
- There are 2 million pairs of items that occur once (and are therefore not frequent) but consist of two frequent items.
- There are 998 million pairs of items that occur once and consist of two infrequent items.

Answer the following questions about an execution of the PCY algorithm on this data:

- (a) On the first pass, how many buckets can we maintain?
- (b) If infrequent pairs distribute as evenly as possible, what is the minimum support threshold for which the PCY algorithm will make a significant reduction in the number of candidate pairs for the second pass (compared with the a-priori algorithm)?

(c) On the assumptions that all frequent pairs are hashed to different buckets, and that infrequent pairs consisting of two frequent items distribute among buckets as evenly as possible, how many candidate pairs are there on the second pass?

(d) How much space is needed on the second pass?

For partial credit, briefly explain your reasoning for each of the parts.

Problem 9 (13 points)

A market-basket data set has 1 million baskets, each of which has 10 items, chosen from among 100,000 different items (not all of which necessarily appear in even one basket). If the support threshold is 100, what are the minimum and maximum numbers of frequent items? What are the minimum and maximum numbers of frequent pairs? For partial credit, briefly explain your reasoning below.

Problem 10 (8 points)

Here is a matrix with three columns and ten rows:

1	0	0
0	1	1
0	1	0
0	0	0
0	0	1
1	1	1
1	0	1
1	1	0
1	1	1
0	0	0

If we choose a random hash function (permutation of the rows), what is the probability that all three columns hash to the same value?

For partial credit, briefly explain your reasoning below.

Problem 11 (8 points)

The method of Alon, Matias, and Szegedy (AMS) that we covered in class was described as a way to compute second moments (surprise number), that is, $\sum m_i^2$, where m_i is the number of occurrences of the i th value in a stream. It can also be used to compute higher order moments. Show that if we want third moments, that is, $\sum m_i^3$, then the proper formula for each variable X is:

1. Pick a random place in the stream, say a place holding value a ,
2. Count the number of occurrences of a from that time forward, say k occurrences.
3. Let the value of X be $n(3k^2 - 3k + 1)$, where n is the length of the stream.

Show that the expected value of X is m_a^3 .

Problem 13 (13 points)

Suppose we are maintaining a count of 1's using the DGIM (buckets) method. Represent a bucket by (i, t) , where i is the number of 1's it represents and t is its timestamp (time of the most recent 1). The current time is 200, the window size is 60, and the current list of buckets is:

$$(16, 148)(8, 162)(8, 177)(4, 183)(2, 192)(1, 197)(1, 200)$$

At the next ten clocks, 201 through 210, the stream has 0101010101. What will the sequence of buckets be at the end of these ten inputs?

Problem 14 (8 points)

Suppose we have some points in a 10-dimensional space, where all points are at the corners of a “cube”; i.e., the only possible points are those that have values 0 or 1 in each dimension. We can thus represent points by bit-strings of length ten. If the distance is the normal Euclidean distance (square root of the sum of the squares of the differences in each dimension), how many points are at distance **less than** 2 from the point 0011010001?

For partial credit, explain your reasoning.

Problem 15 (19 points)

Suppose we want to use the Flajolet-Martin-like approach to estimating the number of distinct values in a stream that we covered in class. There are eight possible values. Imagine that we pick a random perfect hashing function from these values to the bit strings $000, 001, \dots, 111$. That is, each possible hash function assigns each of the eight values to a different bit string. Suppose that three of the eight possible values actually appear in the stream. Let R be the largest number of consecutive 0's at the end of the bit string to which any of those three values maps. A helpful observation is that the number of different hash functions of these three values is $\binom{8}{3}$, or 56. Answer the following questions:

- (a) What is the probability that $R = 3$?

Briefly explain your reasoning.

- (b) What is the probability that $R = 2$?

Briefly explain your reasoning.

- (c) What is the probability that $R = 1$?

Hint: it may be easier to do (d) first and solve this as “everything else.” Briefly explain your reasoning.

(d) What is the probability that $R = 0$?

Briefly explain your reasoning.

(e) What is the expected value of the estimate of the number of different values that comes from any one randomly chosen hash function?

Show your calculation.

(f) Your answer to (e) should not be 3. Give one reason why the F-M method is known to produce an accurate estimate, even though the calculation you just did might be interpreted as suggesting otherwise.