Final 2011 exam paper

Mining Massive Datasets (Stanford University)

CS246 Final Exam Solutions, Winter 2011

1. Your name and student ID.
   - **Name**:..
   - **Student ID**:..
2. I agree to comply with Stanford Honor Code.
   - **Signature**:..
3. There should be XX numbered pages in this exam (including this cover sheet).
4. The exam is open book, open note and open laptop, but you are not allowed to connect to network (3G, WiFi,...). You may use a calculator.
5. If you need more room to work out your answer to a question, use the back of the page and clearly mark on the front of the page if we are to look at what's on the back.
6. Work efficiently. Some questions are easier, some more difficult. Be sure to give yourself time to answer all of the easy ones, and avoid getting bogged down in the more difficult ones before you have answered the easier ones.
7. You have 180 minutes.
8. Good luck!

Question	Topic	Max. score	Score
1	Decision Tree	15	
2	Min-Hash Signature	15	
3	Locality Sensitive Hashing	10	
4	Support Vector Machine	12	
5	Recommendation Systems	15	
6	SVD	8	
7	Map Reduce	16	
8	Advertising	12	
9	Link Analysis	15	
10	Association Rules	10	
11	Similarity Measures	15	
12	K-Means	10	
13	Pagerank	15	
14	Streaming	12 + 10	

1 [15 points] Decision Tree

We have some data about when people go hiking. The data take into effect, wether hike is on a weekend or not, if the weather is rainy or sunny, and if the person will have company during the hike. Find the optimum decision tree for hiking habits, using the training data below. When you split the decision tree at each node, maximize the following quantity:

$$MAX[I(D) - (I(D_L) + I(D_R))]$$

where D, D_L, D_R are parent, left child and right child respectively and $I(D)$ is:

$$I(D) = mH(\frac{m^+}{m}) = mH(\frac{m^-}{m})$$

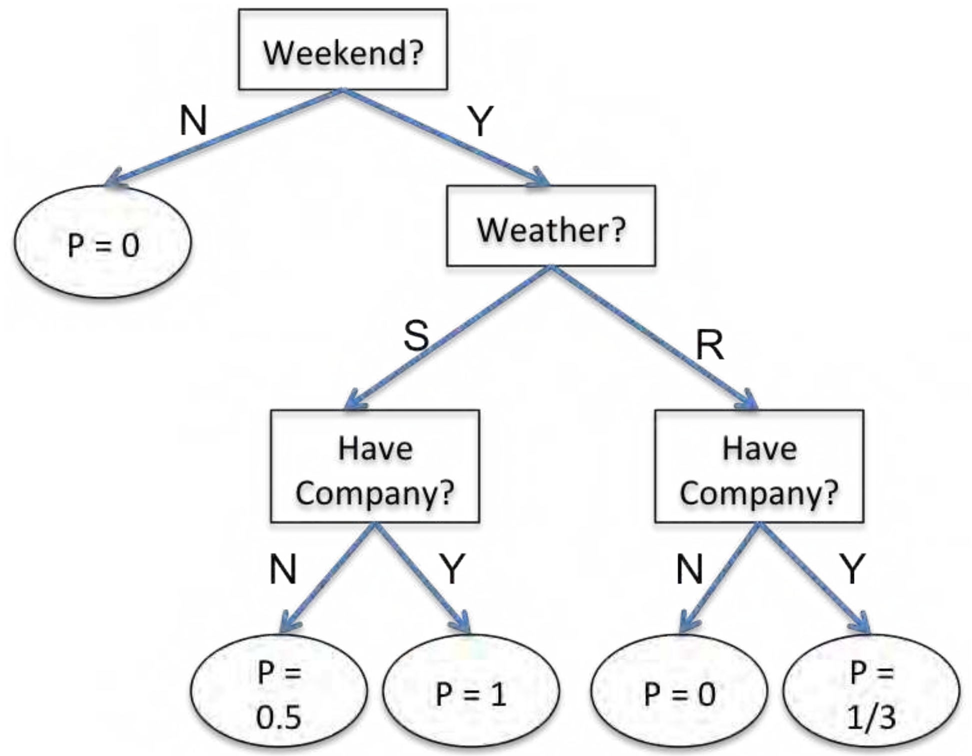
and $H(x) = -x\log_2(x) - (1-x)\log_2(1-x)$, $0 \leq x \leq 1$, is the entropy function and $m = m^+ + m^-$ is the total number of positive and negative training data at the node.

You may find the following useful in your calculations: H(x) = H(1-x), $H(0) = 0$, $H(1/5) = 0.72$, $H(1/4) = 0.8$, $H(1/3) = 0.92$, $H(2/5) = 0.97$, $H(3/7) = 0.99$, $H(0.5) = 1$.

Weekend?	Company?	Weather	Go Hiking?
Y	N	R	N
Y	Y	R	N
Y	Y	R	Y
Y	Y	S	Y
Y	N	S	Y
Y	N	S	N
Y	Y	R	N
Y	Y	S	Y
N	Y	S	N
N	Y	R	N
N	N	S	N

(a) [13 points] Draw your decision tree. **solution:** We want to choose attributes that maximize $mH(p) - m_rH(p_r) - m_lH(p_l)$. This means that at each step, we need to choose the attributes for which $m_rH(p_r) + m_lH(p_l)$ is minimum. For the first step, the *Weekend* attribute achieve this:
$Weekend : m_rH(p_r) + m_lH(p_l) = 8H(1/2) + 3H(0) = 8$
$Weather : m_rH(p_r) + m_lH(p_l) = 5H(1/5) + 6H(1/2) \approx 9.6$
$Company : m_rH(p_r) + m_lH(p_l) = 4H(1/4) + 7H(3/7) \approx 10.1$
Therefore we first split on weekend attribute.
If weekend = NO: then Go Hiking = NO.
If weekend = YES, we need to choose second attribute to split on:
$Weather : m_rH(p_r) + m_lH(p_l) = 4H(1/4) + 4H(1/4) \approx 6.4$
$Company : m_rH(p_r) + m_lH(p_l) = 5H(2/5) + 3H(1/3) \approx 7.6$
Therefore the second attribute will be *Weather* attribute, and third one will be *Company* attribute. The decision tree will be as follows:

(b) [1 point] According to your decision tree, what is the probability of going to hike on a rainy week day, without any company? **Answer:** 0



(c) [1 point] How about probability of going to hike on a rainy weekend when having some company?
Answer: 1/3.

2 [10 points] Min-Hash Signature

We want to compute min-hash signature for two columns, C_1 and C_2 using two psudo-random permutation of columns using the following function:
$h_1(n) = 3n + 2 \bmod 7$
$h_2(n) = n - 1 \bmod 7$
Here, n is the row number in original ordering. Instead of explicitly reordering the columns for each hash function, we use the implementation discussed in class, in which we read each data in a column once in a sequential order, and update the min hash signatures as we pass through them. Complete the steps of the algorithm and give the resulting signatures for C_1 and C_2.

Row	C_1	C_2
0	0	1
1	1	0
2	0	1
3	0	0
4	1	1
5	1	1
6	1	0

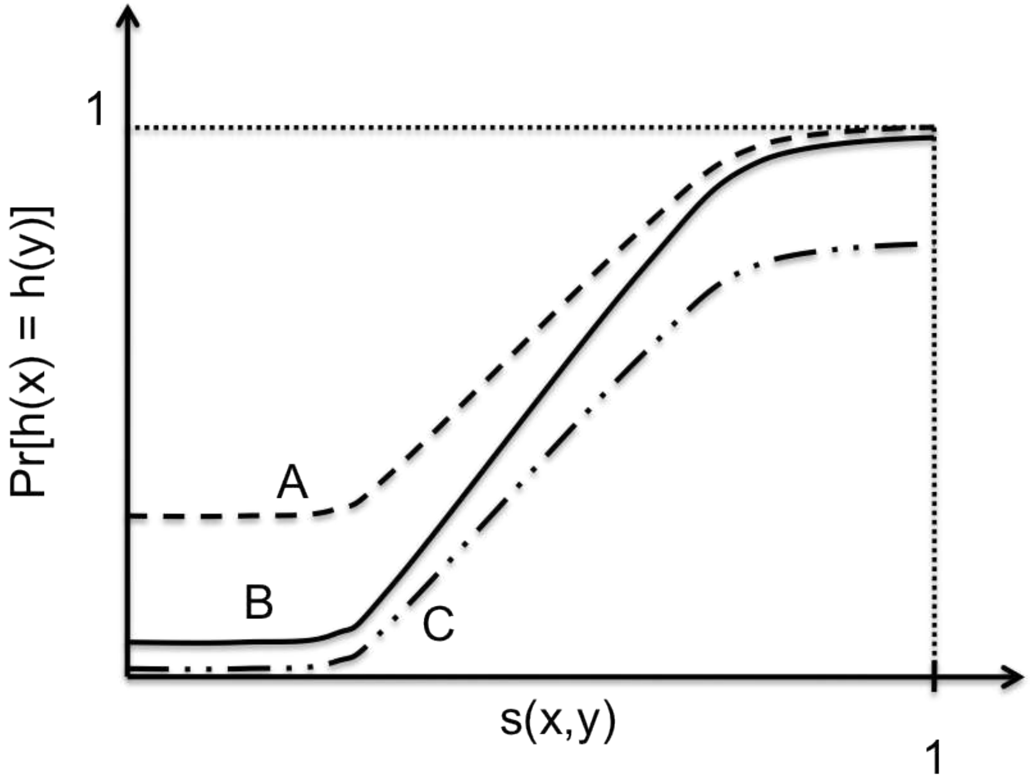
Sig. row#		Sig(C_1)	Sig(C_2)
0	h_1 perm.	-	2
	h_2 perm.	-	6
1	h_1 perm.	5	2
	h_2 perm.	0	6
2	h_1 perm.	5	1
	h_2 perm.	0	1
3	h_1 perm.	5	1
	h_2 perm.	0	1
4	h_1 perm.	0	0
	h_2 perm.	0	1
5	h_1 perm.	0	0
	h_2 perm.	0	1
6	h_1 perm.	0	0
	h_2 perm.	0	1

	Sig(C_1)	Sig(C_2)
h_1 perm.	0	0
h_2 perm.	0	1

3 [10 points] LSH

We have a family of $(d_1, d_2, (1 - d_1), (1 - d_2))$-sensitive hash functions. Using k^4 of these hash functions, we want to amplify the LS-Family using a) k^2-way AND construct followed by k^2-way OR construct, b)k^2-way OR construct followed by k^2-way AND construct, and c) Cascade of a (k, k) AND-OR construct and a (k, k) OR-AND construct, i.e. each of the hash functions in the (k, k) OR-AND construct, itself is a (k, k) AND-OR composition.

Figure below, shows $Pr[h(x) = h(y)]$ vs. the similarity between x and y for these three constructs. In the table below, specify which curve belong to which construct. In one line, justify your answers.

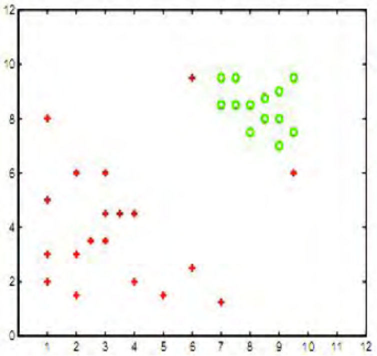


Construct	Curve	Justification
AND-OR	C	AND-OR construct works better in reducing the false positive probability. Hence for small s(x,y), p is the least.
OR-AND	A	OR-AND construct works better in reducing the false negative probability. Hence for large s(x,y), p is the closest to 1.
CASCADE	B	This is getting the best of both worlds.

4 [15 points] SVM

The original SVM proposed was a linear classier. As discussed in problem set 4, In order to make SVM non-linear we map the training data on to a higher dimensional feature space and then use a linear classier in the that space. This mapping can be done with the help of kernel functions.

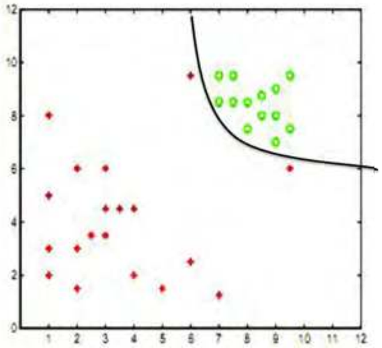
For this question assume that we are training an SVM with a quadratic kernel - i.e. our kernel function is a polynomial kernel of degree 2. This means the resulting decision boundary in the original feature space may be parabolic in nature. The dataset on which we are training is given below:



The slack penalty C will determine the location of the separating parabola. Please answer the following questions qualitatively.

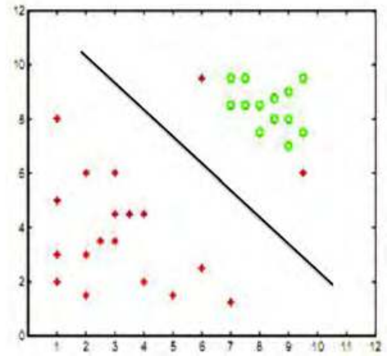
(a) [5 points] Where would the decision boundary be for very large values of C? (Remember that we are using a quadratic kernel). Justify your answer in one sentence and then draw the decision boundary in the figure below.

Answer: Since C is too large, we can't afford any misclassification. Also we want to minimize $\parallel w \parallel$, therefore the x^2 constant is minimum, and hence among all the parabolas, we choose the minimum curvature one.

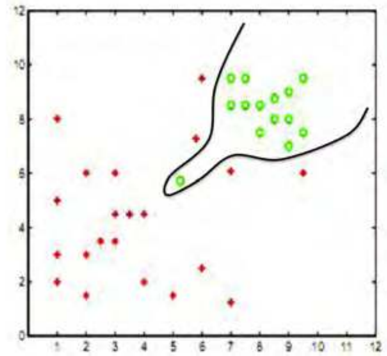


(b) [5 points] Where would the decision boundary be for C nearly equal to 0? Justify your answer in one sentence and then draw the decision boundary in the figure below.

Answer: Since the penalty for mis-classification is too small, the decision boundary will be linear to have x^2 constant equal to 0.



(c) [5 points] Now suppose we add three more data points as shown in figure below. Now the data are not quadratically separable, therefore we decide to use a degree-5 kernel and find the following decision boundary. Most probably, our SVM suffers from a phenomenon which will cause wrong classification of new data points. Name that phenomenon, and in one sentence, explain what it is.



Answer: Over-fitting. It happens when we use an unnecessarily complex model to fit the training data.

5 [10 points] Recommendation Systems

(a) [4 points] You want to design a recommendation system for an online bookstore that has been launched recently. The bookstore has over 1 million book titles, but its rating database has only 10,000 ratings. Which of the following would be a better recommendation system? a) User-user collaborative filtering b) Item-item collaborative filtering c) Content-based recommendation. In One sentence justify your answer.
Answer: *c*. from these choices, the only system that doesn't depend on other users' ratings, is content based recommendation system.

(b) [3 points] Suppose the bookstore is using the recommendation system you suggested above. A customer has only rated two books: "Linear Algebra" and "Differential Equations" and both ratings are 5 out of 5 stars. Which of the following books is less likely to be recommended? a) "Operating Systems" b) "A Tale of Two Cities" c) "Convex Optimization" d) It depends on other users' ratings.
Answer: *b*. In item features, this is probably the furthest from those two books. Note that since the system is content based, choice *d* is not true.

(c) [3 points] After some years, the bookstore has enough ratings that it starts to use a more advanced recommendation system like the one won the Netflix prize. Suppose the mean rating of books is 3.4 stars. Alice, a faithful customer, has rated 350 books and her average rating is 0.4 stars higher than average users' ratings. Animals Farm, is a book title in the bookstore with 250,000 ratings whose average rating is 0.7 higher than global average. What would be a baseline estimate of Alice's rating for Animals Farms?
Answer: $r = 3.4 + 0.7 + 0.4 = 4.5$.

6 [8 points] SVD

(a) [4 points] Let A be a square matrix of full rank, and the SVD of A is given as: $A = U\Sigma V^T$, where U and V are orthogonal matrices. The inverse of A can be computed easily given U, V and Σ. Write down an expression for A^{-1} in their terms. Simplify as much as possible. **Answer:**

ANSWER: Answer: $A^{-1} = V\Sigma^{-1}U^T$

(b) [4 points] Let us say we use the SVD to decompose a Users × Movies matrix M and then use it for prediction after reducing the dimensionality. Let the matrix have k singular values. Let the matrix M_i be the matrix obtained after reducing the dimensionality to i singular values. As a function of i, plot how you think the error on using M_i instead of M for prediction purposes will vary.

ANSWER: Answer: Will reduce then increase

7 [16 points] MapReduce

Compute the total communication between the mappers and the reducers (i.e., the total number of (key, value) pairs that are sent to the reducers) for each of the following problems: (Assume that there is no combiner.)
(a) [4 points] Word count for a data set of total size D (i.e., this is the total number of words in the data set.), and number of distinct words is w. **Answer:**

ANSWER: Answer: D
(b) [6 points] Matrix multiplication of two matrices, one of size $i \times j$ the other of size $j \times k$ in one map-reduce step, with each reducer computing the value of a single (a, b) (where $a \in [1, i], b \in [1, k]$) element in the matrix product. **Answer:**

ANSWER: Answer: $2ijk$: Intuition is that we send each item in the first matrix to all k corresponding to the second matrix, and each item in the second matrix to all i in the first matrix.
(c) [6 points] Cross product of two sets — one set A of size a and one set B of size b ($b \ll a$), with each reducer handling all the items in the cross product corresponding to a single item $\in A$. As an example, the cross product of two sets $A = \{1, 2\}, B = \{a, b\}$ is $\{(1, a), (1, b), (2, a), (2, b)\}$. So there is one reducer generating $\{(1, a), (1, b)\}$ and the other generating $\{(2, a), (2, b)\}$. **Answer:**

ANSWER: Answer: $a \times b + a$

8 [12 points] Advertising

Suppose we apply the BALANCE algorithm with bids of 0 or 1 only, to a situation where advertiser A bids on query words x and y, while advertiser B bids on query words x and z. Both have a budget of $2. Identify the sequences of queries that will certainly be handled optimally by the algorithm, and provide a one line explanation.

(a) yzyy **Answer:**

ANSWER: Answer: YES. Note that the optimum only yields $3.

(b) xyzx **Answer:**

ANSWER: Answer: YES. Whichever advertiser is assigned the first x, the other will be assigned the second x, thus using all four queries.

(c) yyxx **Answer:**

ANSWER: Answer: YES. The two x's will be assigned to B, whose budget is the larger of the two, after the two y's are assigned to A.

(d) xyyy **Answer:**

ANSWER: Answer: NO. If the first x query is assigned to A, then the yield is only $2, while $3 is optimum.

(e) xyyz **Answer:**

ANSWER: Answer: NO. If the x is assigned to A, then the second y cannot be satisfied, while the optimum assigns all four queries.
(f) xyxz **Answer:**

ANSWER: Answer: NO. Both x's could be assigned to B, in which case the z cannot be satisfied. However, the optimum assigns all four queries.

9 [15 points] Link Analysis

Suppose you are given the the following topic-sensitive page-rank vectors computed on web graph G, but you are not allowed to access the graph itself.

- r1, with teleport set $\{1, 2, 3\}$
- r2, with teleport set $\{3, 4, 5\}$
- r3, with teleport set $\{1, 4, 5\}$
- r4, 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) [5 points] r5, corresponding to the teleport set $\{2\}$ **Answer:**

ANSWER: Answer: YES. 3r1 - 3r2 + 3r3 - 2r4

(b) [5 points] r6 with teleport set $\{5\}$ **Answer:**

ANSWER: Answer: NO. Cannot distinguish between 4/5

(c) [5 points] r7, with teleport set $\{1, 2, 3, 4, 5\}$, with weights 0.1,0.2,0.3,0.2,0.2 respectively. **Answer:**

ANSWER: Answer: YES. 0.3 (2r1 + r2 + r3) - 0.2 r4

10 [10 points] Association Rules

Suppose our market-basket data consists of $n(n-1)/2$ baskets, each with exactly two items, There are exactly n items, and each pair of items appears in exactly one basket. Note that therefore each item appears in exactly $n-1$ baskets. Let the support threshold be $s = n-1$, so every item is frequent, but no pair is frequent (assuming $n > 2$). We wish to run the PCY algorithm on this data, and we have a hash function h that maps pairs of items to b buckets, in such a way that each bucket gets the same number of pairs.

a) [5 points] Under what condition involving b and n will there be no frequent buckets? **Answer:**

ANSWER: Answer:

$$\frac{n(n-1)}{2b} < n-1$$

b) [5 points] If all counts (i.e., counts of items and the counts for each bucket) require 4 bytes, how much memory do we need to run PCY in main memory? Your answer should be a function of n only. **Answer:**

ANSWER: Answer: $6n + 4$ bytes.

11 [15 points] Similarity Measures

In class we discussed the Jaccard similarity of columns of a boolean matrix. We used letters a, b, c, and d to stand, respectively, for the numbers of rows in which two columns had 11, 10, 01, and 00, respectively, and we determined that the Jaccard similarity of the columns was a/a+b+c. An alternative measure of similarity for columns is the Hamming similarity, which is the fraction of the rows in which these columns agree. Let j(x,y) and h(x,y) be, respectively, the Jaccard and Hamming similarities of columns x and y.

(a) [4 points] In terms of a, b, c, and d, give a formula for the Hamming similarity of columns. **Answer:** (a+d)/(a+b+c+d)

(b) [6 point] Indicate if each of the statements below is true or false. If true, show it with an example; if false, give a one sentence explanation why it is false.

(1) h(x,y) can be greater than j(x,y) **Answer:** True. $x = (0, 1)$, $y = (0, 0)$.

(2) h(x,y) can be equal to j(x,y) **Answer:** True. $x = (1, 1)$, $y = (1, 1)$.

(3) h(x,y) can be less than j(x,y) **Answer:** False. For h(x,y) to be less than j(x,y), we need a>a+b+c, which is impossible.

(c) [5 point] The Hamming and Jaccard similarities do not always produce the same decision about which of two pairs of columns is more similar. Your task is to demonstrate this point by finding four columns u, v, x, and y (which you should write as row-vectors), with the properties that j(x,y) > j(u,v), but h(x,y) < h(u,v). Make sure you report the values of j(x,y), j(u,v), h(x,y), and h(u,v) as well. **Answer:** $u = (0, 1, 0, 0)$, $v = (0, 0, 0, 1)$, $x = (1, 0, 1, 0)$, $y = (1, 1, 0, 1)$. Then, j(u,v)=0, j(x,y)=h(x,y)= 0.25, h(u,v)= 0.5

12 [10 points] K-means

With a dataset $\mathcal{X}$ to be partitioned into k clusters, recall that the initialization step of the k-means algorithm chooses an arbitrary set of k centers $\mathcal{C} = \{c_1, c_2, \ldots, c_k\}$. We studied two initialization schemes, namely random and weighted initialization methods. Now, consider the following initialization method which we denote as the "Greedy" initialization method, which picks the first center at random from the dataset, and then iteratively picks the datapoint that is furthest from all the previous centers. More exactly:

1. Choose c_1 uniformly at random from $\mathcal{X}$
2. Choose the next center c_i to be $\text{argmax}_{x \in \mathcal{X}}\{D(x)\}$.
3. Repeat step 2 until k centers are chosen.

where at any given time, with the current set of cluster centers $\mathcal{C}$, $D(x) = \min_{c \in \mathcal{C}} ||x - c||$.

With an example show that with the greedy initialization, the k-means algorithm may converge to a clustering that has an arbitrarily larger cost than the optimal clustering (i.e., the one with the optimal cost). That is, given an arbitrary number $r > 1$, give an example where k-means with greedy initialization converges to a clustering whose cost is at least r times larger than the cost of the optimal clustering. Remember that the cost of a k-means clustering was defined as:

$$\phi = \sum_{x \in \mathcal{X}} \min_{c \in \mathcal{C}} ||x - c||^2$$

Answer: Assume $k = 2$, and consider a dataset in the two dimensional plane, formed of m points at $(-1, 0)$, m points at $(0, 1)$, and one point at $(0, d)$ for some $d >> 1$ (e.g. $d = 100$). The clustering with centers $(-1, 0)$ and $(0, -1)$ has cost d^2, but the greedy initialization picks $(0, d)$ as one of the centers, and hence, as $m \to \infty$, gets an arbitrarily worse clustering.

13 [15 points] PageRank

Consider a directed graph $G = (V, E)$ with $V = \{1, 2, 3, 4, 5\}$, and $E = \{(1, 2), (1, 3), (2, 1), (2, 3), (3, 4), (3, 5), (4, 5), (5, 4)\}$.

(a) [5 points] Set up the equations to compute pagerank for G. Assume that the "tax" rate (i.e., the probability of teleport) is 0.2. **Answer:** Denoting the pagerank of node i with $\pi(i)$ $(1 \leq i \leq 5)$, we have:

$$\begin{aligned}
\pi(1) &= 0.04 + 0.4\pi(2) \\
\pi(2) &= 0.04 + 0.4\pi(1) \\
\pi(3) &= 0.04 + 0.4\pi(1) + 0.4\pi(2) \\
\pi(4) &= 0.04 + 0.4\pi(3) + 0.8\pi(5) \\
\pi(5) &= 0.04 + 0.4\pi(3) + 0.8\pi(4)
\end{aligned}$$

(b) [5 point] Set up the equations for topic-specec pagerank for the same graph, with teleport set $\{1, 2\}$. Solve the equations and compute the rank vector. **Answer:** Denoting the topic-specific pagerank of node i with $\pi'(i)$ $(1 \leq i \leq 5)$, we have:

$$\begin{aligned}
\pi'(1) &= 0.1 + 0.4\pi'(2) \\
\pi'(2) &= 0.1 + 0.4\pi'(1) \\
\pi'(3) &= 0.4\pi'(1) + 0.4\pi'(2) \\
\pi'(4) &= 0.4\pi'(3) + 0.8\pi'(5) \\
\pi'(5) &= 0.4\pi'(3) + 0.8\pi'(4)
\end{aligned}$$

(c) [5 point] Give 5 examples of pairs (S, v), where $S \subseteq V$ and $v \in V$, such that the topic-specific pagerank of v for the teleport set S is equal to 0. Explain why these values are equal to 0. **Answer:** $\{(3, 1), (3, 2), (4, 1), (4, 2), (5, 1), (5, 2)\}$. For node v to have a non-zero topic-specific pagerank with teleport set S, we need at least one directed path from a node in S to v. But, there is no such path from 3 to 1, and so on.

14 [12 points + 10 extra points] Streaming

Assume we have a data stream of elements from the universal set $\{1, 2, \ldots, m\}$. We pick m independent random numbers r_i $(1 \leq i \leq m)$, such that:

$$Pr[r_i = 1] = Pr[r_i = -1] = \frac{1}{2}$$

We incrementally compute a random variable Z: At the beginning of the stream Z is set to 0, and as each new element arrives in the stream, if the element is equal to j (for some $1 \leq j \leq m$), we update: $Z \leftarrow Z + r_j$.

At the end of the stream, we compute $Y = Z^2$.

(a) [12 points] Compute the expectation $E[Y]$. **Answer:** If x_j $(1 \leq j \leq m)$, is the number of times j arrives as an element of the stream, then $Z = \sum_{j=1}^{m} r_j x_j$, and hence $E[Y] = E[Z^2] = E[\sum_{1\leq i,j \leq m} r_i r_j x_i x_j] = \sum_{1\leq i,j\leq m} E[r_i r_j] x_i x_j$. But, $E[r_i r_j] = 1\{i = j\}$, hence $E[Y] = \sum_{i=1}^{m} x_i^2$.

(b) [EXTRA CREDIT 10 points] (ONLY ATTEMPT WHEN DONE WITH EVERYTHING ELSE!) Part (a) shows that Y can be used to approximate the surprise number of the stream. However, one can see that Y has a large variance. Suggest an alternative distribution for the random variables r_i such that the resulting random variable Y has the same expectation (as in part (a)) but a smaller variance. You don't need to formally show that the variance of your suggested estimator is lower, but you need to give an intuitive argument for it. **Answer:** As long as $E[r_i] = 0$, $E[r_i^2] = 1$, and r_i's are picked independently, the proof in part (a) still goes through. So, to decrease the variance, we only need to pick r_i's from a distribution with the mentioned properties, which also has a light tail. The Gaussian $N(0, 1)$ distribution is one such candidate.