CS246: Mining Massive Data Sets **Spring 2023**

Exam

- Your Name: _____
- Your SUNetID: _____
- Your SUID: _____

I acknowledge and accept the Stanford Honor Code.

Signature: _____

1. There are 13 questions in this exam (numbered 2 to 14); the maximum score that you can obtain is 180 points. These questions require thought but do not require long answers. Please be as concise as possible. You can use the number of points as a rough estimate of how long we think a question may take you. And don't worry, we don't expect you to finish all the questions.
2. This exam is open-book and open-notes. You may use notes (digitally created notes are allowed) and/or lecture slides and/or any reference material. However, **answers should be written in your own words**.
3. Acceptable uses of computer:
   - You may access the Internet, but you may not communicate with any other person. Similarly, AI-driven code completion tools including ChatGPT and GitHub Copilot are not allowed
   - You may use your computer to write code or do any scientific computation, though writing code is not required to solve any of the problems in this exam.
   - You can use your computer as a calculator or an e-reader.
4. Collaboration with other students is not allowed in any form. Please do not discuss the exam with anyone until after grades are released.
5. If you have any clarifying questions, make a **private post on Ed**. It is very important that your post is private; if it is public, we may deduct points from your exam grade.
6. Please submit your answers here on Gradescope. You have three options to submit your answers: (1) to upload one file for all questions (at the top of the Gradescope exam) (2) to upload one file per question, in a file upload field in the last sub-question; or (3) to write your answers directly in the text fields in the sub-questions.
7. Numerical answers may be left as fractions, as decimals rounded to **2 decimal places**, or as radicals (e.g., $\sqrt{2}$).

8. The exam will be graded on a curve. Do not stress if you are unable to finish every question in time.

9. **Please save your responses in a separate document as Gradescope does not automatically save partially completed submissions**

2 True/False Questions (15 points)

For each of the statements given below, mark them as either True or False. For full credit, also add a 1-2 line reasoning behind your answer.

1. **(1.5 point)** If ABC is a frequent itemset and BCD is not a frequent itemset, then ABD cannot be a frequent itemset

2. **(1.5 point)** Increasing the number of hash tables in LSH decreases the probability of false positives.

3. **(1.5 point)** K-means clustering can handle outliers well by assigning them to their own clusters.

4. **(1.5 point)** Similar to k-means clustering, hierarchical clustering methods also require a predefined number of clusters.

5. **(1.5 point)** The TF-IDF score of a term in a document is proportional to the frequency of the term in the document and inversely proportional to the number of documents that mention the term.

6. **(1.5 point)** A group of bloggers decides to cross-promote by linking each other's websites on all of their blogs. This is a spider trap because if we represent each website as a node, the bloggers' websites will form a loop.

7. **(1.5 point)** Embeddings learned through neural networks are always linearly separable in the embedding space.

8. **(1.5 point)** The information gain at the root node of the decision tree is greater than or equal to the information gain at any other node.

9. **(1.5 point)** Following the terminology of the row sampling method for matrix sketching, let's consider a set of items $a_1, \ldots, a_n$ with corresponding weights $w_1, \ldots, w_n$. Our goal is to obtain a small and representative sample. As per the lectures, we define a representative sample as one that accurately estimates the total weight of the selected items in expectation. In this context, allowing higher weight items to be sampled with higher probability increases bias (the difference from actual total weight in expectation) but reduces variance.

10. **(1.5 point)** The greedy algorithm has a higher competitive ratio than generalized version of BALANCE algorithm.

3 Locality Sensitive Hashing (16 points)

You would like to use shingling and locality sensitive hashing to identify possible plagiarism in student essays. There are two methods you would like to try:

- Compare an essay P with a publication Q in your database using shingling and Jaccard similarity.
- Measure similarity with cosine similarity, where the shingle sets are viewed as binary vectors

In the following questions, consider an essay P with 2000 unique length-c shingles in it and a publication Q with 6000 unique length-c shingles, let S_P and S_Q be the shingle sets for P and Q respectively. So $|S_P| = 2000$ and $|S_Q| = 6000$. Assume the essay P was fully copied from a portion of Q. Note that the shingle size c is arbitrary and will not factor into any of the solutions below.

1. **(6 points)** What is the Jaccard similarity between the shingle sets of the two documents? Also calculate $\Pr(\text{MinHash}(S_P) = \text{MinHash}(S_Q))$.
2. **(8 points)** Let m be the number of all possible length-c shingles (i.e., all ordered sets of c length words in the English language). Represent S_P and S_Q by length m binary vectors x_P and x_Q, with a 1 in every position corresponding to a shingle they contain.
   - What is the cosine similarity between x_P and x_Q (again assuming P was fully copied from a portion Q)?
   - What is $\Pr(\text{SimHash}(x_P) = \text{SimHash}(x_Q))$?
     Hint: Use that for any two vectors z, w, $\Pr(\text{SimHash}(z) = \text{SimHash}(w)) = 1 - \frac{\theta}{\pi}$ where θ is the angle between z and w in radians.
3. **(3 points)** Given the above, which similarity metric and hash function would you pick for the plagiarism detection task?

4 Clustering (12 points)

1. **(8 points)** We would like to perform hierarchical clustering on the following dataset of n points – $2^0, 2^1, \cdots, 2^n$. Let's say we use the euclidean distance and our stopping condition is when there is only 1 cluster left. Provide a visual representation of the hierarchical clustering trees (dendrogram) resulting from the following linkage strategies:

   (a) single linkage - merge clusters based on the closest pairwise distance between points.

   (b) complete linkage - merge clusters based on the farthest pairwise distance between points.

   (c) average linkage - merge clusters based on the average pairwise distance between points.

2. **(4 points)** Now we modify our distance function as follows:

$$d(x, y) = \frac{\max(x, y)}{\min(x, y)}$$

   where x, y are any two 1-D points. For the same dataset as in previous part, construct the hierarchical clustering trees for complete linkage method. For this part, assume that the total number of points satisfy $n = 2^k$ for some integer k.

   In case of ties, use the following strategy - define rank of a point 2^i as $i + 1$ and rank of a cluster is the minimum rank of points within the cluster. Now to break ties, first merge the clusters with the smallest rank.

5 Dimensionality Reduction (12 points)

Consider the following 3×3 matrix: $A = \begin{pmatrix} 3 & 1 & 0 \\ 1 & 2 & -1 \\ 0 & -1 & 3 \end{pmatrix}$.

Singular Value Decomposition. You are given the decomposition $A = SDS^{-1}$ with the diagonal matrix D. (Note that SVD usually breaks A into UDV^T but for a symmetric matrix (like A), the SVD decomposition becomes $A = SDS^{-1}$).

$$A = SDS^{-1} = \begin{pmatrix} \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{3}} & 0 & \frac{2}{\sqrt{6}} \\ -\frac{1}{\sqrt{3}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} \end{pmatrix} \begin{pmatrix} 4 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{6}} & \frac{2}{\sqrt{6}} & \frac{1}{\sqrt{6}} \end{pmatrix}$$

1. **(3 points)** Write an expression for A^k for any positive integer k. You can write it in terms of S and D where S and D are from the SVD decomposition of A. Show your calculations for full credit.

2. **(4 points)** Write an expression for e^A in terms of S and D, where e^A is called a matrix exponential and is defined by the expression:

   $$e^A = \sum_{k=0}^{\infty} \frac{1}{k!} A^k$$

   . Show your calculations for full credit.

   Hint: Exponential is similarly defined for a real number:

   $$e^x = \sum_{k=0}^{\infty} \frac{1}{k!} x^k$$

   .

3. **(5 points) Low-rank Approximation**. Let N be the rank 2 approximation of M in terms of reconstruction Frobenius norm error. Remember that N is also be a 3×3 matrix. Calculate N and its singular value decomposition. (Note: Frobenius norm of a matrix A is defined as follows: $||A_F|| = \sqrt{\Sigma(A_{ij}^2)}$)

6 Recommender Systems (12 points)

Suppose you have a database of books, with information about their genre, author, and year of publication. You also know which users have read and liked each book (rating 1) or disliked it (rating 0).

The table below summarizes the database:

Book	Year	Genre	Author	Total number of ratings
F	2012	Mystery	H1	100
G	2020	Mystery	H1	1000
H	2016	Horror	H2	600
I	2004	Horror	H2	40
J	2014	Comedy	H3	1

Consider user U1, who is interested in the year 2016, the author H2, and the genre Comedy. You have a recommender system R that suggested the book G to user U1. R could be one or more of the following options:

- User-user collaborative filtering
- Item-item collaborative filtering
- Content-based recommender system

Answer the following questions:

1. **(4 points)** Which option(s) do you think R could be? Explain your answer. (If more than one option is possible, you need to state them all.)
2. **(4 points)** Independent of the previous sub-part, assume R is user-user collaborative filtering. How can R recommend a book to a new user U2? Under what conditions can it do so? If it cannot, why not?
3. **(4 points)** Item-item collaborative filtering is seen to work better than user-user because users have multiple tastes. But this also means that users like to be recommended a variety of movies.

   Given the genre of each book (there are 3 different genres in the dataset) and an item-item collaborative filtering recommender system that predicts k top-movies to a user (k can be an input to the recommender), suggest a way to find top 5 movies to a user such that the recommender will try to incorporate books from different genres as well. (Note: Explain in 3–5 lines maximum, no rigorous proof is required.)

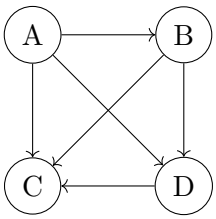
7 PageRank (14 points)

1. **(3 points)** How can teleportation be useful in personalized PageRank?

2. **(3 points)** Why do we only consider incoming edges when calculating the importance of a node?

3. **(4 points)** Construct a graph of 5 nodes where each node has the same PageRank no matter what the teleportation probability is set to. Justify your answer in 1-2 lines.

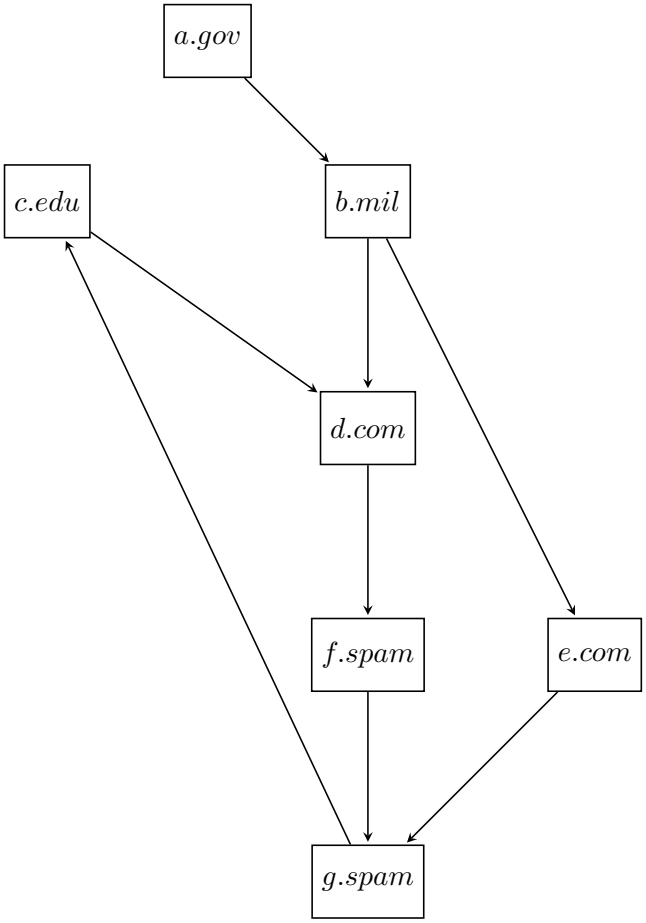
4. **(4 points)** What is the column-stochastic adjacency matrix M corresponding to the below graph?



8 Extended PageRank (15 points)

In your first week as a new intern at a search engine startup, you are tasked with handling spam links.

1. **(7 points)** Your manager tells you to use the Trusted Propagation model to deal with spam where each website node will have a trust value associated with it and trusts below a threshold will be eliminated as spam before we calculate PageRank. Your manager wants you to use the trusted domain strategy where the trust of .gov, .edu and .mil pages are set to 1. Given a trust threshold of 0.08, what is the maximum value of β that will ensure that ALL spam nodes in the below graph are flagged correctly? Show your work.



2. **(4 points)** Your coworker tells you that spam pages cannot get high page ranks and picking the top k pages as the seed set is better than using trusted domains. For the above graph, would this approach work well? Justify your answer.

3. **(4 points)** The startup decides to pivot to building an image recommendation system. Your manager tells you to copy Pixie, Pinterest's recommendation algorithm and to minimize execution costs as much as possible. You know that Pixie uses early stopping to reduce the number of steps in random walk simulations. What could happen if you set a low value for the minimum threshold of visit counts to warrant stopping early?
   Hint: Consider what will happen if you make an algorithm simulate random walks on a

bipartite graph with more than 1000 nodes and a density greater than 5 and stop until the 100th most visited pin/node has at least 2 visits.

9 Community Detection (12 points)

In your first week as a new intern at the social media company MiceBase, your manager tasks you with devising a less computationally intensive method for performing community detection, given a set of seed nodes. From what you can tell, the specified graph of user interactions seems to be clearly split into densely connected communities which are themselves sparsely connected to each other. You decide to collect advice from your trusted friends and coworkers before bringing your findings back to your manager. For each of the following claims, evaluate whether the advice is correct. If the answer is incorrect, briefly explain why. If the answer is correct, briefly discuss whether it is useful to your task.

1. **(4 points)** Your coworker Eliezer claims that Personalized PageRank can be faithfully computed without accessing the entire graph by instead running the algorithm on the induced subgraph formed by randomly sampling a subset of nodes uniformly at random with likelihood p, as long as $p \geq 1 - \frac{1}{e}$.

2. **(4 points)** Your friend John claims that we can significantly improve the runtime of the Personalized PageRank algorithm by first partitioning the graph into N disconnected subgraphs, each containing exactly one of the N communities that is present in the full graph. He argues that since the clusters in the graph are very dense, it will be easy to partition the graph in this way.

3. **(4 points)** You reach out to one of your trusted sources online, and they send the following reply: "As an AI language model I am not able to provide authoritative advice on what solution is appropriate for your situation, but it's important to keep in mind that K-NN graph methods will generally be less effective than approximate PPR methods. This is because the K-NN graphs are randomly constructed, so they don't carry useful information about the structure of your graph."

10 Learning Embeddings (14 points)

1. **(3 points)** You are tasked with creating a movie recommender system using a dataset that contains thousands of movies. In this context, you need to decide which specific model, CBOW (Continuous Bag-of-Words) or Skip-gram, would be most appropriate for feature representation of the movie documents. Provide an explanation for your choice, considering the characteristics of each model and their suitability for the movie recommender system.

2. **(7.5 points)** For the word2vec algorithm, select all correct answers below:

   (a) Word2vec takes as input a document-word matrix.

   (b) Word2vec models local structure between a center word and neighboring words up to a fixed distance in the input text sequence.

   (c) Word2vec models can be computed using a deep network with L2 output loss.

   (d) Word2vec models can be computed using a deep network with cross-entropy output loss.

   (e) Word2vec computes two word embedding matrices U and V for center and context words respectively.

3. **(3.5 points)** One known problem with word embeddings is that antonyms, which are words with opposite meanings, often have similar embeddings. For example, by searching for most similar words to "increase" or "enter" using cosine similarity, words with clear antonyms like "decrease" or "exit" are likely to appear among the top similar words. Discuss why embeddings might have this tendency.

11 Decision Trees (18 points)

Our goal is to create a decision tree for n training examples (x, y) where each point x consists of d features or attributes.

1. **(3 points)** Let's consider a scenario where there are indices i and j such that for all examples x in our training data, the features $x^{(i)}$ and $x^{(j)}$ are equal. Additionally, in cases where the conditional entropy is equal, we prioritize $x^{(i)}$ as the tiebreaker. If we were to remove feature j from our training data, would it have any impact on the decision tree learned for this dataset? Please provide a brief explanation.

2. **(3 points)** Let's consider a scenario in which our dataset includes two identical training points p and q, that is, we have $x_p = x_q$ and $y_p = y_q$. If we remove point q from the dataset, would it impact the learned decision tree? Please provide a brief explanation.

3. **(2 points)** Let's consider a scenario where we intend to utilize binary splits on attribute/feature $x^{(1)}$ which can take one out of n values. Splitting involves selecting a value (denoted as a) to partition the data such that $x^{(1)} < a; x^{(1)} \geq a$ are the two splits. In this case, we aim to ensure that each branch of the split contains at least one vector. What is the total number of values of a that we need to consider?

4. **(3 points)** Let's consider a situation where we aim to employ three-way splits. In order to achieve this, we must select two values (a, b such that $a < b$) and then partition the data into three sets based on the following criteria: $x^{(1)} < a; a \leq x^{(1)} < b; x^{(1)} \geq b$. Similar to the previous part, we want to ensure that each branch of the split contains at least one vector for any considered values of a and b. How many $\{a, b\}$ pairs do we need to consider?

5. **(7 points)** Is it possible to recreate a three-way split at the root (parameterized by $\{a, b\}$) using a tree that only uses binary splits? Justify your answer.

   If yes, explain the potential benefits of learning a three-way split tree. If no, explain which type of data can be accurately classified by three-way splits but not with binary splits.

12 Mining Data Streams (12 points)

Suppose you are working for a social media company called Teeter. The company collects a stream of user posts on various topics, such as sports, politics, entertainment, etc. You are given the task of counting how many distinct users have posted on politically extreme topics. The company policy is very strict towards politically extreme topics. They are okay with counting some politically non-extreme posting users but **NOT** okay with missing any politically extreme posting users.

Formally, you are given a stream of <user,topic> tuples and a set P of politically extreme topics. You know that the total number of users on the platform is $|U| = 100,000$.

Given your CS246 experience, you decide to apply a combination of Bloom Filters and Flajolet-Martin algorithm. Your new algorithm works as follows:

1. Pick a bloom filter on topics using a bit array of $|B|$ bits and k different hash functions.
2. Use the bloom filter to select tuples from the stream whose topic belongs to the set P of politically extreme topics, and remove the remaining tuples from the stream.
3. To the stream of selected tuples, apply the Flajolet-Martin algorithm over the users to count the number of unique users.

Based on the above algorithm, answer the following questions:

1. (4 points) You are first experimenting with the bloom filter to find the optimal k value which minimizes the false positive rate. So far, you have the rough measurements given in the following table.

k	False Positive Rate
4	0.91%
7	0.37%
11	0.55%
13	1.22%

   To conduct further experiments, which of the following range(s) of values should you focus your efforts on.

   (a) $k < 4$
   (b) $4 < k < 7$
   (c) $7 < k < 11$
   (d) $11 < k < 13$
   (e) $13 < k$

2. (4 points) Let's assume you figured out the optimal k and that gives you a false positive rate of 0.14%. Now you apply Flajolet-Martin algorithm. You note that the maximum number of trailing zeros in the binary strings you have seen so far is 14. Calculate the estimated number of politically extreme posting users and show your calculations. You don't have to calculate the final value, feel free to leave the answer in expression form. (*Hint: Don't forget to take the bloom filtering into account*).

3. (4 points) Suddenly, you hear an announcement that your company Teeter has been acquired by Enol Dusk. Thankfully, you haven't been fired. However, the company policy on politically extreme posts has changed from very strict to very relaxed. Now, counting politically moderate posting users is **NOT** okay but missing some politically extreme posting users is fine. Does the algorithm you designed still work? Explain your reasoning.

13 Matrix Sketching (10 points)

Consider the following 4×4 matrix A filled with small random integer values:

$$A = \begin{bmatrix} 2 & 3 & 0 & 4 \\ -1 & 0 & 2 & 1 \\ 3 & 2 & -2 & 0 \\ 1 & 1 & 4 & 2 \end{bmatrix}$$

We will apply the CUR method with $k = 2$, where k represents the number of columns/rows to be sampled. To determine the C and R matrices, we will treat the norm of each row or column as their weight and sample the top 2 norm columns or rows.

The norm of each column/row: $\text{norm}(c_i)$ and $\text{norm}(r_i)$, where c_i represents a column and r_i represents a row $(1 \leq i \leq 4)$, is given as follows:

$$\text{norm}(c_1) = \sqrt{2^2 + (-1)^2 + 3^2 + 1^2} = \sqrt{15} \approx 3.87$$

$$\text{norm}(c_2) = \sqrt{3^2 + 0^2 + 2^2 + 1^2} = \sqrt{14} \approx 3.74$$

$$\text{norm}(c_3) = \sqrt{0^2 + 2^2 + (-2)^2 + 4^2} = \sqrt{24} = 4.89$$

$$\text{norm}(c_4) = \sqrt{4^2 + 1^2 + 0^2 + 2^2} = \sqrt{21} \approx 4.58$$

$$\text{norm}(r_1) = \sqrt{2^2 + 3^2 + 0^2 + 4^2} = \sqrt{29} \approx 5.38$$

$$\text{norm}(r_2) = \sqrt{(-1)^2 + 0^2 + 2^2 + 1^2} = \sqrt{6} \approx 2.45$$

$$\text{norm}(r_3) = \sqrt{3^2 + 2^2 + (-2)^2 + 0^2} = \sqrt{17} \approx 4.12$$

$$\text{norm}(r_4) = \sqrt{1^2 + 1^2 + 4^2 + 2^2} = \sqrt{22} \approx 4.69$$

1. **(1.5 point)**Sample and mention the top 2 norm columns to form the C matrix.
2. **(1.5 point)** Sample and mention the top 2 norm rows to form the R matrix.
3. **(4 points)** Now, using the obtained C and R matrices, compute the U matrix by taking the pseudo-inverse of intersection of the sampled columns and rows.

   You can use https://www.omnicalculator.com/math/pseudoinverse for calculating the pseudo-inverse of a matrix.

4. **(3 points)** Finally, calculate the reconstructed matrix.

   You can use https://www.wolframalpha.com/input/?i=matrix+multiplication+calculator for calculating the matrix multiplication of 2 (or 3) matrices.

14 Computational Advertising (18 points)

In this question, we will consider a new algorithm for computational advertising called k-Greedy-BALANCE. Before we explain the problem, please note that we will restrict the input space to the following constraints:

1. There are three types of queries: x, y, and z.
2. There are two ads: 1. Ad A, which can be shown for queries x and y, and 2. Ad C, which can be shown for queries x and z.
3. Both ads have a budget of $\$B$ and the same bid amount of \$1 for all types of queries.
4. The total number of queries is $2B$.
5. Only configurations are allowed where the optimal advertising solution achieves $\$2B$ revenue.

Now, let's describe the k-Greedy-BALANCE algorithm:

1. Read the query.
2. If only one of the ads A or C can be served for the query, serve it.
3. If both ads A and C can be served and have unspent budget:
   - Serve ad C if the unspent budget of C is at least $\$k$ more than the unspent budget of A (i.e., unspent_budget$(C) \geq$ unspent_budget $+ \$k$).
   - Otherwise, serve ad A.

For the following questions, we will assume $k = \frac{B}{5}$ (assume B is a multiple of 5). Remember that all competitive ratios are calculated using the input space defined under the before mentioned constraints. Also remember from the lectures that within our assumptions, the competitive ratio of the Greedy algorithm is $\frac{1}{2}$, and the competitive ratio of the BALANCE algorithm is $\frac{3}{4}$. You are free to use any content or result proven in the lectures as is, just remember to cite that the result/content has been proven/used in lectures.

1. **(5 points)** Show that the competitive ratio of k-Greedy-BALANCE is less than $\frac{3}{4}$. (Hint: You can demonstrate a counterexample by showing a query sequence for which the optimal solution achieves $2B$ revenue, but k-Greedy-BALANCE achieves less than $\frac{3B}{2}$.)
2. **(5 points)** Show that the competitive ratio of k-Greedy-BALANCE is greater than $\frac{1}{2}$. (Hint: Think about why the worst-case scenario of the normal Greedy algorithm wouldn't work here.)
3. **(8 points)** Show that the competitive ratio of k-Greedy-BALANCE within our input constraints is $\frac{7}{10}$.