Test 2: Streaming, Dimensionality Reduction, Noise, Graphs

—— Practice ——

NAME: UID: FINAL SCORE:

This real test will allow one 8×11.5 inch notes sheet, front and back. Electronic devices that can transmit/receive information will **not** be allowed (e.g., computers, phones, calculators, ipads). Unlimited blank scratch paper is allowed.

Absolutely no talking will be allowed, unless a TA or instructor is present and you are asking a question. Talking students will have their tests confiscated.

Show your work to have an increased chance for partial credit.

1 Outliers (10 points)

A: (10 points) Let $X \in \mathbb{R}^2$ be some data you want to find k clusters from using distance function $D : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+$.
Describe a way to find outliers.
Draw an pictorial example of how this might work.

2 Streaming (25 points)

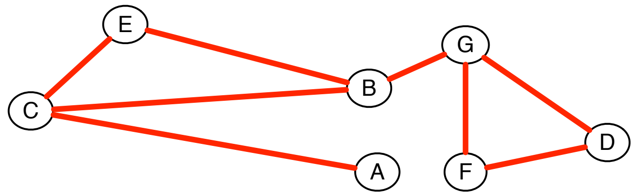
A: (10 points) Describe why algorithms we discussed in class for betweenness of a graph, for PageRank using the power iteration, and Distance Metric Learning are difficult to operate in a stream.
*(I am looking for a very short discussion about the streaming model – what is it – and what aspect of these algorithms do not align with that model. You do **not** need a separate answer for each algorithm.)*

B: (15 points) Consider the Misra-Gries algorithm with 3 counters and labels. Consider part way through the stream it has the following state:

labels	a	c	w
counters	32	1	7

Let the next 3 elements of the stream be $\langle a, b, g \rangle$. Show the state of the 3 labels and counters of the Misra-Gries algorithm after processing each of those items.

3 Graphs (25 points)



Consider the pictured graph.

A: (5 points) Which edge has the largest betweenness score?

B: (20 points) Consider creating a Markov chain from the graph's adjacency matrix, in the standard way we showed in class. *[Each undirected edge (e.g. $\{A, C\}$) is treated as two directed edges (e.g. (A, C) and (C, A)). The the transition probability matrix is defined over all vertices, where the next state is chosen as the vertex pointed to by one of the outgoing edges, chosen uniformly at random. For instance, if you are at state F, then with probability $1/2$ the next state is D and with probability $1/2$ the next state is G.]*

1. Write down the probability transition matrix.

2. Is this Markov Chain ergodic? Explain why or why not.

4 Dimensionality Reduction (40 points)

Consider a single data point $a = (10, 0, 0, -2)$ in 4 dimensions.

Let $u_1 = (\frac{1}{2}, 0, -\frac{1}{\sqrt{2}}, \frac{1}{2})$ and $u_2 = (-\frac{1}{2}, \frac{1}{\sqrt{2}}, 0, \frac{1}{2})$ be basis vectors.

A: (10 points) Which of a, u_1 and u_2 are unit vectors? Explain why.

B: (5 points) Are u_1 and u_2 orthogonal? Explain why.

C: (15 points) Show how to project a to a two-dimensional space spanned by u_1 and u_2. That is, write the two coordinates of a in this new two-dimensional space (show your work).

D: (5 points) Calculate the L_2 norm of the new 2-dimensional vector (show your work).

E: (5 points) Can the L_2 norm of vector ever be larger after a projection using two orthogonal unit vectors? If so, provide an example.