CMPSCI 711: HOMEWORK
DUE 5PM, TUESDAY MAY 1ST

- You should attempt any seven of the following questions (although feel free to attempt more). Solutions should be typed. You may work in groups of at most two and hand in one copy of your solutions per group.

Question 1. Given a vector $f \in \mathbb{R}^n$, here's an alternative sketch to estimate $F_2 = \|f\|_2^2$. Consider d four-wise independent hash functions $h_1, \ldots h_d$ from $[n] \to [w]$ and compute $c_{r,i} = \sum_{j:h_r(j)=i} f_j$. For each $r \in [d]$, let

$$e_r = \sum_{i=1}^{w/2} (c_{r,2i} - c_{r,2i-1})^2 \ .$$

The estimate for F_2 is the median of $e_1, e_2, \ldots, e_d$. Compute the mean and variance of each e_r and show that if $w = O(\epsilon^{-2})$ and $d = O(\log 1/\delta)$, we obtain a $1+\epsilon$ estimate of F_2 with probability at least $1-\delta$.

Question 2. This question is about estimating the size of the minimum vertex cover in a tree.

(1) Let T be a tree with at least 3 nodes. Let k be the number of nodes with degree at least 2 and let c be the size of the minimum vertex cover. Prove the best upper and lower bound you can on the quantity c/k.

(2) Given a arbitrarily ordered stream of edges that define a tree on n nodes, design a randomized data stream algorithm using $O(\text{poly}(1/\epsilon)\text{polylog}n)$ space that returns a $2+\epsilon$ approximation of the size of the minimum vertex cover of this tree. **Hint:** Consider using an F_0 (the number of non-zero entries) estimation algorithm and a vector d where d_i is the degree of the ith node. See `https://people.cs.umass.edu/~mcgregor/711S18/vectors-2.pdf` for a description of a sketch for F_0 estimation but you don't actually need to know how the sketch works, just that it's a linear sketch.

Question 3. This problem is on pairwise independence.

(1) Let $X_1, X_2, \ldots, X_k$ be independent binary random variables where $Pr[X_i = 0] = Pr[X_i = 1] = 1/2$. Given $S \subseteq \{1, 2, \ldots, k\}, S \neq \emptyset$ define a random variables $Y_S = \bigoplus_{i \in S} X_i$. Prove that the collection of random variables $\{Y_S | S \subseteq \{1, 2, \ldots, k\}, S \neq \emptyset\}$ are pairwise independent.

(2) We say a family of hash functions $\mathcal{H} = \{h_1, h_2, \ldots\}$ from set A to set B is 2-universal family if, for any $x, y \in A$ when we pick $h \in \mathcal{H}$ uniformly at random then

$$\Pr(h(x) = h(y)) \leq 1/|B| \ .$$

Suppose $A = [n]$ and $B = [n^3]$. Prove that the probability that a random h from $\mathcal{H}$ is injective is at least $1 - 1/n$.

Question 4. Given a graph $G = (V, E)$ with n nodes and m edges, the density $d(U)$ of a subset of nodes $U \subseteq V$ is defined as the number of edges whose endpoints are both in U divided by $|U|$. Let maximum density of G is defined as $d^* = \max_{U \subseteq V} d(U)$. Suppose we sample each edge independently with probability p and let E' be the set of edges sampled. Prove that if $p = \min(1, c\epsilon^{-2}(\log n)n/m)$ for some large c, then d^* can be estimated up to a factor $1+\epsilon$ from $G' = (V, E')$ with high probability. **Hint:** First argue that $d^* \geq m/n$ and then use some combination of the Chernoff Bound and the Union Bound.

Question 5. Consider a code $x_1, x_2, \ldots, x_n \in [q]^m$ with minimum distance

$$d = \min_{i \neq j} |\{k \in [m] : x_i[k] \neq x_j[k]\} .$$

Let $h_i : [n] \to [q]$ be the hash function defined by $h_i(j) = x_j[i]$, i.e., the ith symbol of the j codeword. Consider the sketch of the vector $f \in \mathbb{R}^n$ where the entries of the sketch are

$$c_{i,j} = \sum_{k:h_i(k)=j} f_k \qquad \forall i \in [m], j \in [q]$$

Given this sketch, what's is the best error bound (in terms of d and m) you can prove for point queries. You may assume all entries of f are non-negative. If you are familiar with results in coding theory, translate this into a bound in terms of n and a user-defined parameter ϵ such that the error is linear in ϵ.

Question 6. Suppose Alice knows $x \in \{0,1\}^n$, Bob knows $y \in \{0,1\}^n$, and they want to compute

$$f(x,y) = \begin{cases} 1 & \text{if } x = y \\ 0 & \text{if } x \neq y \end{cases} .$$

Prove a lower bound on the number of bits that need to be communicated between Alice and Bob in any deterministic protocol even if the protocol is allowed an unlimited number of rounds. For simplicity you may assume that the players take turns sending just one bit. We can prove the bound as follows: Consider the $2^n \times 2^n$ matrix M where rows are indexed by possible strings Alice could have, columns are indexed by possible strings Bob could have, and $M[x', y'] = f(x', y')$ for each pair $x', y' \in \{0,1\}^n$. Define a rectangle to be any subset of $\{0,1\}^n \times \{0,1\}^n$ that can be expressed as $A \times B$ where $A \subseteq \{0,1\}^n$ and $B \subseteq \{0,1\}^n$. Prove by induction that for any $c \geq 0$,

$$\{(x,y) \in \{0,1\}^n \times \{0,1\}^n : (x,y) \text{ is consistent with the first } c \text{ bits communicated so far}\}$$

is a rectangle. Next, argue that any protocol in which the maximum number of bits communicated is h, partitions $\{0,1\}^n \times \{0,1\}^n$ into at most 2^h rectangles. If the protocol is always correct, f should take the same value for any (x,y) pair in the same rectangle. Use this fact to argue that h needs to be at least n.

Question 7. Consider a stream of m edges that describe a graph $G = (V, E)$ on n nodes. We say a set of 3 nodes $\{u, v, w\}$ forms a triangle if there is an edge between every pair of these nodes. Let T be the total number of triangles in G and suppose we have a guarantee that $T \geq t$ for some given value t. Consider the following basic algorithm for estimate T: pick an edge $\{u, v\}$ uniformly at random from the stream; pick a vertex w uniformly at random from $V \setminus \{u, v\}$; output $m(n-2)$ if edges $\{u, w\}$ and $\{v, w\}$ occur after $\{u, v\}$ in the stream, and 0 otherwise. Prove that the output of this algorithm has expectation exactly T. Suppose we run the basic algorithm multiple times in parallel and average the output. How many times to we need to run the basic algorithm such that the final answer is within a factor of $(1 + \epsilon)$ of T with probability at least $1 - \delta$.

Question 8. Consider a graph stream describing an unweighted, undirected n-vertex graph G. Prove that $\Omega(n^2)$ space is required to determine, in one pass, whether or not G contains a triangle, even with randomization allowed and the algorithm may fail with probability at most $1/3$.

Question 9. Consider a stream of m values in the range $[n]$ and let f_i be the number of times i occurs in the stream. The Count-Min Sketch solves the problem of estimating any f_j, given j, but does not directly give us a quick way to identify, e.g., the set of elements with frequency greater than some threshold. In this question, we fix this.

In greater detail: Let α be a constant with $0 < \alpha < 1$. We would like to maintain a suitable summary of the stream (it will be some extension of Count-Min Sketch) so that we can, on demand,

quickly produce a set $S \subseteq [n]$ satisfying the following properties with high probability: (1) S contains every j such that $f_j \geq \alpha m$; (2) S does not contain any j such that $f_j < (\alpha - \epsilon)m$. Design a data stream algorithm that achieves this. Your space usage, as well as the time taken to process each token and to produce the set S, should be polynomial in the parameters, $\log m, \log n, \log(1/\delta)$ and $1/\epsilon$, and may depend arbitrarily on α.

Question 10. Suppose a stream consists of a sequence of edges and their weights. Design a small-space deterministic stream algorithm that constructs a data structure that can be used to $(1+\epsilon)(2t-1)$-approximate any distance in the graph.