HW 4

Due on **Monday, April 4, 2022 at 10am**

CS 574: Randomized Algorithms, Spring 2022 Version: **1.11**

Submission guidelines same as previous homework.

10 (100 PTS.) Expander graph.

Let $\varepsilon \in (0,1)$ be a fixed constant. Prove that for any n sufficiently large, there is a constant degree graph G with n vertices, such that for any set $B \subseteq \mathsf{V}(\mathsf{G})$ with $|B| \leq n/2$, the graph $\mathsf{G} - B$ is almost connected, where $\mathsf{G} - B = \big(\mathsf{V}(\mathsf{G}) \setminus B, \{uv \in \mathsf{E}(\mathsf{G}) \mid u, v \in \mathsf{V}(\mathsf{G}) \setminus B\}\big)$. Formally, there is a connected component in $\mathsf{G} - B$ that contains at least $(1-\varepsilon)(n - |B|)$ vertices. What is the maximum degree of the graph as a function of ε? You solution needs to be self contained, and use only techniques seen in class.

(Hint: Look on the construction seen in class for expanding graphs. You want the maximum degree of the constructed graph to be as small as possible as a function of ε.)

(The question becomes somewhat easier if the maximum degree is allowed to be poly-logarithmic. Such a solution would get almost all the points for this question.)

11 (100 PTS.) Streaming inverse frequencies.

Let S be a stream of m numbers taken out of the set $\{1, \ldots, n\}$. For a number $i = 1, \ldots, n$, let f_i be the number of times that i appears in the stream. You can assume that this stream contains at least $m/10$ distinct values.

You are also given parameter $\varepsilon \in (0,1)$ and $\delta \in (0,1)$. Design an algorithm that uses small space (in both n and m) that outputs an estimate H to the quantity $G = \sum_{i : f_i \neq 0} \frac{1}{f_i}$. The algorithm is required to have the property that

$$\mathbf{Pr}[(1-\varepsilon)G \leq H \leq (1+\varepsilon)G] \geq 1 - \delta.$$

Provide full details!

(Hint: Follow the algorithms seen in class. Figure out where you have to use the provided assumption.)

12 (100 PTS.) Estimate it.

Let $S = s_1, s_2, \ldots, s_m$ be a stream (m is not known to you in advance). Given an item s_i in the stream, you can check if it is valid by calling an oracle D. Let τ be the number of valid elements in the stream. The oracle calls are expensive. Given parameters ε, δ, describe an algorithm that outputs an estimate u for the number of items in the stream that are valid, such that $\mathbf{Pr}[\tau - \varepsilon m \leq u \leq \tau + \varepsilon m] \geq 1 - \delta$.

Your algorithm needs to use little space, and few oracle calls. In particular, how many oracle calls does your algorithm performs, say in expectation. How much space does your algorithm uses?

13 (100 PTS.) Streaming under deletion.

Assume you that can use the JL lemma, but you do not have to pay for the space used by the lemma (magic!). You are given a stream $S \equiv s_1, \ldots, s_m$ of operations, each operation is either an insertion or deletion of element out of a set $\{1, \ldots, n\}$. At time t in the stream, and a number k, its frequency $f(i,t) = \text{ins}(k,t) - \text{del}(k,t)$, where $\text{ins}(k,t)$ is the number of times k was inserted in the first t operations, and $\text{del}(k,t)$ is how many times it was deleted.

At any time during the streaming, the algorithm might stop, and ask for $1 \pm \varepsilon$-estimate for $\sum_{i=1}^{n} \bigl(f(i,t)\bigr)^2$. The estimate should be correct with probability $\geq 1 - n^{-10}$.

Show how to do this using little space. (Hint: Maintain implicitly the vector $(f(1,t), \ldots, f(n,t))$.)