Fall 2014, CS 598: Algorithms for Big Data

Homework 3
Due: 10/23/2014

Instructions and Policy: Each student should write up their own solutions independently. You need to indicate the names of the people you discussed a problem with; ideally you should discuss with no more than two other people.

Solve as many problems as you can.

You need to typeset your solutions in Latex. Please write clearly and concisely - clarity and brevity will be rewarded. Refer to known facts as necessary. Your job is to convince me that you know the solution, as quickly as possible.

Problem 1 In a turnstile stream updating a vector $x \in \mathbb{R}^n$ starting as the 0 vector, an ϵ-*error* ℓ_1 *sampler* is a streaming algorithm that when queried outputs a pair $(i, \hat{x}_i)$ such that i is output with probability $|x_i|/\|x\|_1$ and $\hat{x}_i = (1 \pm \epsilon)x_i$ (recall that in turnstile streams, each stream update is of the form $x_i \leftarrow x_i + v$ where v can be positive or negative). Pretend we have such an ℓ_1 sampler using space $S(n, \epsilon)$. Now consider the following problem: you see a stream $i_1\ i_2\ \cdots\ i_{n+1}$ with each $i_j \in [n]$. This stream must have at least one duplicate entry due to the pigeonhole principle. Show how to use a 1/2-error ℓ_1 sampler to give a one-pass streaming algorithm that reports at least one duplicate index $i \in [n]$ with probability at least $1 - \delta$. The space of your algorithm should be $O(S(n, 1/2) \cdot \log(1/\delta))$.

Problem 2 We have seen streaming algorithms for ϵ-approximate quantiles. We defined An ϵ-approximate quantile for a quantile $\phi \in (0, 1]$ as an element of rank r where $\phi n - \epsilon n \leq r \leq \phi n + \epsilon n$ where n is number of elements. We define a stronger notion of ϵ-approximate quantiles where we wish to return an element of rank r where $(1 - \epsilon)\phi n \leq r \leq (1 + \epsilon)\phi n$. Describe how to compute an ϵ-approximate quantile summary for this stronger notion of approximation.

Problem 3 We saw an algorithm in the semi-streaming model for finding a constant factor approximation to the maximum weight matching problem. Now consider the following variant. We are given a graph $G = (V, E)$ and each edge e has a weight $w(e)$. Moreover each edge has a color from $\{1, 2, \ldots, k\}$ and each color i has a bound $b(i)$. The goal is to find a maximum weight matching M which satisifes the additional constraint that the number of edges in M from a color class i is at most $b(i)$. Assume that you are given the $b(i)$'s ahead of time and the each edge when it arrives in the stream specifies its end points, its weight and its color. Describe a constant factor approximation for this problem in the semi-streaming setting.