Problem Set 9

Randomized Algorithms

Due Tuesday, November 16

1. In this problem we develop a locality sensitive hash for Jaccard similarity of documents. Let W be the set of all possible words.

   Given two sets $A, B \subset W$, the *Jaccard similarity* of A and B is

   $$J(A,B) = \frac{|A \cap B|}{|A \cup B|}.$$

   Given two text documents, we can measure the similarity between them as the Jaccard similarity between the set of words in each document (the bag-of-words model).

   (a) Suppose you are given a uniformly random function $f : W \to [0,1]$ from words to the unit interval. Let $h : \mathcal{D} \to W$ be the "MinHash" function:

   $$h(A) = \arg\min_{w \in A} f(w).$$

   Show that

   $$\Pr[h(A) = h(B)] = J(A,B).$$

   (b) Suppose you are given a constant $C > 1$ and parameter $r \in (0, 1/C)$. For any constant $\epsilon > 0$, show how to construct a hash function g such that:

      i. For any two sets A, B with $J(A,B) < 1 - Cr$, $\Pr[g(A) = g(B)] = 1/n$.

      ii. For any two sets A, B with $J(A,B) > 1 - r$, $\Pr[g(A) = g(B)] > n^{-1/C-\epsilon}$.

2. In class we showed that network coding works well on a static graph. The key property was that, if vertex v is "aware" of a vector u in one round, then each neighbor becomes aware of it in the next round independently with probability at least $1-1/q$. We showed that this implies that after R rounds, the destination t becomes aware of each u with probability $1-q^{-C_{s,t}R(1-\epsilon)}$, where $C_{s,t}$ is the (s,t) min cut and suitably large parameters ($q > 2^{O(1/\epsilon)}$ and $R > O(n/\epsilon)$).

   In this problem we extend this to dynamic graphs. We instead suppose that the graph changes arbitrarily in every round, with the condition that the (s,t) min cut is at least C in each round.

   (a) For any u that s is aware of, at the beginning of round i let S_i be the set of vertices that are aware of u. Show that, if $t \notin S_i$, then over the randomness in round i we have
   $$\Pr[S_{i+1} = S_i] \leq q^{-C}.$$
   That is to say, almost always at least one new vertex will become aware of u.

   (b) Show that, after $R \geq O(n/\epsilon)$ rounds and with $q \geq 2^{O(1/\epsilon)}$,
   $$\Pr[t \notin S_R] \leq q^{-CR(1-\epsilon)}.$$